

ORIGINAL COPY

AAMRL-SR-90-507

AD-A231 401



**SPEECH PERCEPTION AND PRODUCTION IN
SEVERE ENVIRONMENTS**

DAVID B. PISONI

INDIANA UNIVERSITY
BLOOMINGTON IN 47402

SEPTEMBER 1990

FINAL REPORT FOR THE PERIOD JULY 1986 TO JUNE 1990

DTIC
ELECTE
JAN 29 1991
S B D

Approved for public release; distribution is unlimited

HARRY G. ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY
HUMAN SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-6573

91 1 25 092

NOTICES

When US Government drawings, specifications or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Federal Government agencies registered with Defense Technical Information Center should direct requests for copies of this report to:

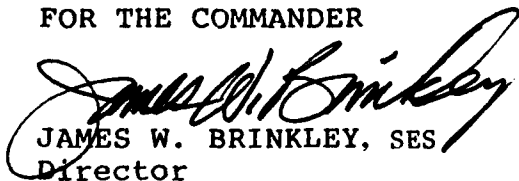
Defense Technical Information Center
Cameron Station
Alexandria, Virginia 22314

TECHNICAL REVIEW AND APPROVAL

AAMRL-SR-90-507

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER



JAMES W. BRINKLEY, SES
Director

Biodynamics and Bioengineering Division
Armstrong Aerospace Medical Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to: Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE September 1990	3. REPORT TYPE AND DATES COVERED Final Report 1 Jul 86 - 3 Jun 90	
4. TITLE AND SUBTITLE Speech Perception and Production in Severe Environments			5. FUNDING NUMBERS PR-7231 TA-32 WU-06 C - F33615-86-C-0549 PE-61102F	
6. AUTHOR(S) David B. Pisoni				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Indiana University P.O. Box 1847 Bloomington IN 47402			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) HARRY G. ARMSTRONG AEROSPACE MEDICAL RESEARCH LABORATORY WRIGHT-PATTERSON AFB OH 45433-6573			10. SPONSORING/MONITORING AGENCY REPORT NUMBER AAMRL-SR-90-507	
11. SUPPLEMENTARY NOTES AAMRL/Contact: Dr Thomas J. Moore, AAMRL/BBA Telephone (513) 255-3607				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Unclassified/unlimited			12b. DISTRIBUTION CODE A	
13. ABSTRACT (Maximum 200 words) The goal of this project was to acquire new knowledge about speech perception and production in severe environments such as high masking noise, increased cognitive load or sustained attentional demands. We examined changes in speech production under these adverse conditions through acoustic analysis techniques. One set of studies focused on the effects of noise on speech production. The experiments in this group were designed to generate a database of speech obtained in noise and in the quiet. A second set of experiments was designed to examine the effects of cognitive load on the acoustic-phonetic properties of speech. Talkers were required to carry out a demanding perceptual motor task while they read lists of test words. A final set of experiments explored the effects of vocal fatigue on the acoustic-phonetic properties of speech. Both cognitive load and vocal fatigue are present in many applications where speech recognition technology is used, yet their influence on speech production is poorly understood.				
14. SUBJECT TERMS Acoustic Phonetics. Speech Production Stress Effects			15. NUMBER OF PAGES	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UNLIMITED	

Final Report

Air Force Contract No. F33615-86-C-0549

"Speech Perception and Production in Severe Environments"

7/1/86 - 6/3/90

DoD Contractor:

Department of the Air Force
Air Force Systems Command
Aeronautical Systems Div/PMRNB
Wright-Patterson AFB, Ohio 45433

Contract Monitor:

Thomas J. Moore, Ph.D.
AFAMRL/BBA
Wright-Patterson AFB, Ohio 45433

Institution

Indiana University
P.O. Box 1847
Bloomington, Indiana 47402
(812) 855-0516

Principal Investigator:

David B. Pisoni, Ph.D.
Professor of Psychology
Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405
(812) 855-1155


David B. Pisoni, Principal Investigator

Submission Date: August 29, 1990

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or
A-1	Specimen

Table of Contents

Overall Goals and Objectives	3
Project I. Speech Production and Perception in Noise	3
Project II. Effects of Cognitive Load on Speech Production	9
Project III. Effects of Vocal Fatigue	14
Summary and General Conclusions	15
Staffing	17
Contract Related Travel	17
Publications	18
Conference Presentations and Invited Talks	18
SRL Progress Report and Technical Note Series	18
References	19
Appendices	23

Perception and Production of Speech in Severe Environments
AAMRL Contract No. F33615-86-C-0549
7/1/86 - 6/30/90

Overall Goals and Objectives

The primary goal of this research project was to acquire new basic knowledge about speech perception and production in severe environments such as high masking noise, increased cognitive load or sustained attentional demands. We examined changes in speech production under these adverse conditions through acoustic analysis techniques. We were also concerned with the listener who may be required to make rapid responses to spoken messages that are physically degraded.

One set of studies focused on the effects of noise on speech production. The experiments in this group were designed to generate a database of speech obtained in noise and in the quiet. A second set of experiments was designed to examine the effects of cognitive load on the acoustic-phonetic properties of speech. Talkers were required to carry out a demanding perceptual motor task while they read lists of test words. A final set of experiments explored the effects of vocal fatigue on the acoustic-phonetic properties of speech. Both cognitive load and vocal fatigue are present in many applications where speech recognition technology is used, yet their influence on speech production is poorly understood. Our general goal in these studies was to learn more about speech perception and production under severe environmental conditions and to acquire new fundamental knowledge about the underlying mechanisms that talkers employ in their speech to compensate for these factors.

Taken together, the research obtained from these three sets of studies has provided a better understanding of how speech is produced and perceived under a variety of demanding conditions and how talkers modify the way they speak. Such fundamental knowledge should contribute to improved designs and applications of speech communication technologies in human-to-human communication and speech I/O technologies in human-to-machine communication. The use of speech I/O technologies in highly demanding environments should reduce operator workload and ultimately improve performance in a variety of applications.

Project I. Speech Production and Perception in Noise

For many years, it has been known that talkers will produce speech with greater intensity when they are required to speak in noisy environments. The effect was first described by Lombard in 1911. Lane and colleagues have offered the hypothesis that increases in vocal amplitude occur so that the speaker can maintain constant intelligibility (Lane, Tranel, and Sisson, 1970; Lane and Tranel, 1971). Little work has been done, however, to quantify the changes that occur at the acoustic-phonetic level under these conditions. The first study we carried out assessed these changes using a variety of test vocabularies. We collected production data from several subjects and performed speech intelligibility tests with the utterances produced in noise.

Air Force Vocabulary

We have completed acoustic analyses and perceptual experiments using utterances from two talkers who reproduced the Air Force vocabulary under several conditions of masking noise. These utterances were produced in the quiet and in 80, 90, and 100 dB of masking noise. The acoustic analyses demonstrated that increases in ambient noise produced changes in word amplitude, word duration, fundamental frequency, and spectral tilt (the relative distribution of energy across frequencies). We have also examined the F1 and F2 frequencies of these utterances. Our analyses of the formant data suggest that speech produced in noise contains higher F1 frequencies than speech produced in the quiet. This effect was, however, subject to individual variation. Figure 1 displays mean F1 frequency data collapsed across utterances for each of the two speakers.

Insert Figure 1 about here

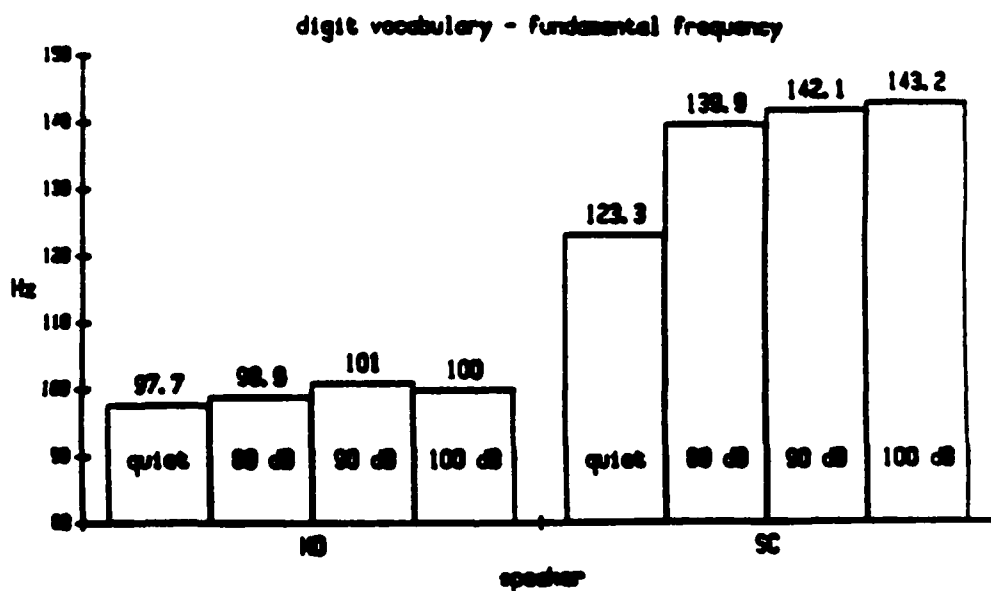
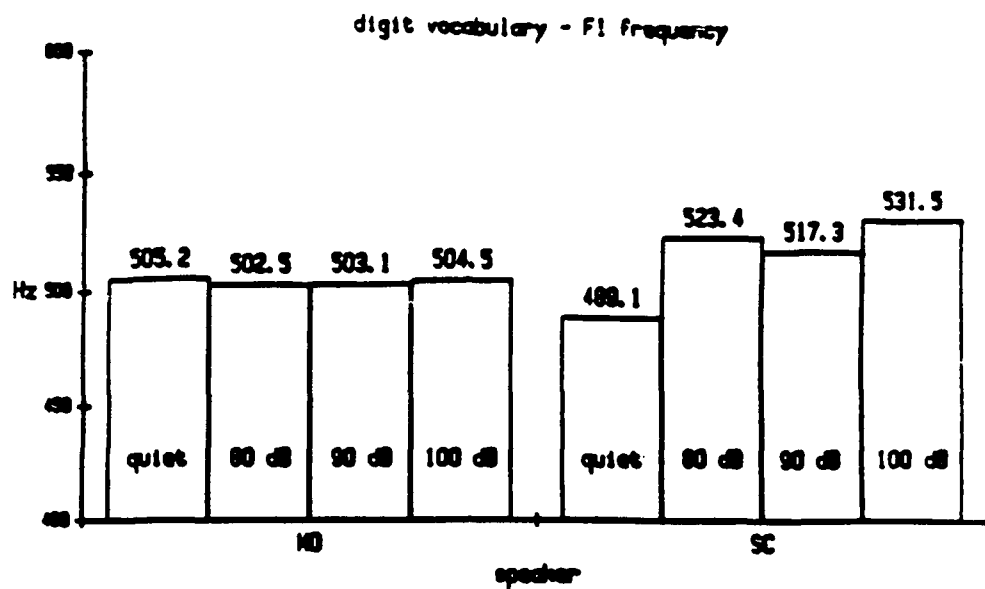
For speaker MD, there was little change in the mean frequency of F1 across noise conditions. However, for speaker SC, F1 was consistently higher for utterances produced in masking noise compared to utterances produced in the quiet. For each speaker, the overall pattern of results for F1 was similar to the pattern found in the analysis of F0 (see Figure 2). That is, for speaker MD, F0 showed little change across noise conditions, whereas for speaker SC, F0 values were higher for utterances produced in the masking noise conditions than in the quiet condition. Our initial interpretation of these findings was that the increases in F1 found for speaker SC were directly related to increases in F0. However, further analyses, in which F0 was treated as a covariate, established that the increases in F1 for utterances produced in noise were not dependent upon increases in F0.

Insert Figure 2 about here

Turning to the results for F2, we found little change in F2 frequencies across noise conditions for either speaker. For speaker SC, however, there was a tendency for the range of F2 frequencies to be reduced for utterances produced in masking noise. That is, utterances containing high F2 frequencies showed a decrease in F2 when produced in noise and utterances containing low F2 frequencies showed an increase in F2.

Perceptual Intelligibility Experiments with the Air Force Vocabulary

We conducted two perceptual experiments to determine whether the observed changes in the spectral and temporal properties of speech produced in noise might influence speech intelligibility. In each experiment, utterances produced in the quiet and utterances from one of the masking noise conditions were presented to a panel of listeners for identification. All utterances were equated for RMS amplitude and mixed with white noise at a constant S/N ratio. Separate groups of listeners were tested at S/N ratios of -5, -10, and -15 dB. In the first experiment, subjects identified MD's and SC's productions of digits that were selected from the quiet and 90 dB noise



Figures 1 and 2. Mean values for first formant frequency and fundamental frequency for talkers MD and SC at four ambient noise level (digit vocabulary).

conditions. In the second experiment, utterances from the quiet and 100 dB noise condition were used. The results of the two experiments are shown in Figures 3 and 4.

Insert Figures 3 and 4 about here

In each experiment, speech produced in noise was more intelligible than speech produced in the quiet. The pattern was consistent for both speakers at every S/N ratio tested. Analyses of variance on the data from each experiment verified that utterances produced in 90 dB and 100 dB of masking noise were significantly more intelligible than utterances produced in the quiet ($p < .0001$). In addition, we found significant interactions between masking noise and S/N ratio in each experiment ($p < .05$). These interactions reflected the fact that differences in intelligibility between speech produced in the quiet and in noise tended to increase as S/N ratio decreased.

Talking in Noise: /h/-vowel-/d/ Vocabulary

We also carried out acoustic analyses to examine the influence of masking noise on speech production using more phonetically controlled vocabularies than the original Air Force vocabulary, which consisted of the digits one through nine and several control words. The first of these studies used an /h/-vowel-/d/ vocabulary in which the 10 English monophthongs i, I, ~~f~~, A, O, ae, a, U, u, ~~e~~ were used as medial vowels. Three speakers produced multiple tokens of each of these hVd utterances in the quiet and in 90 dB of masking noise.

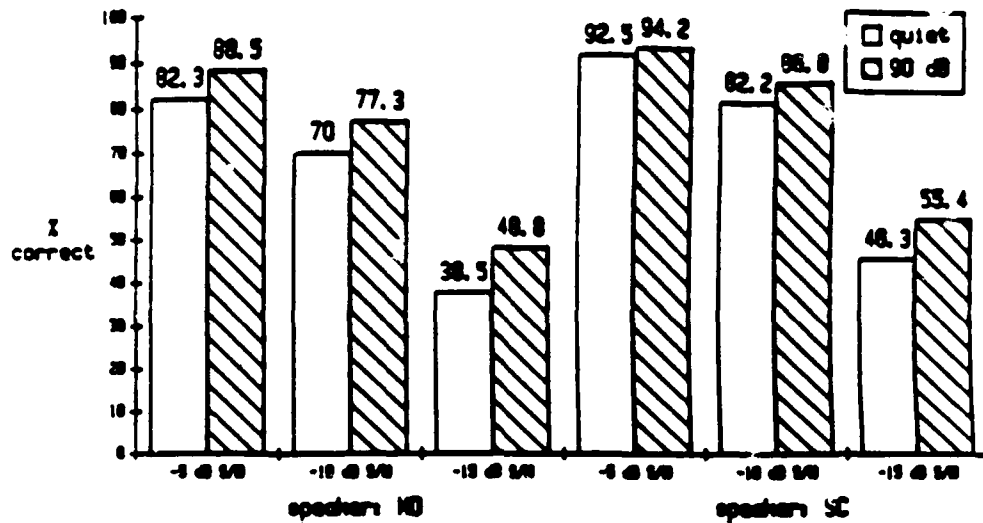
The hVd study allowed us to assess the reliability of our previous findings in terms of the influence of environmental noise on amplitude, pitch, duration, and spectral tilt. Figures 5 through 8 present these results: mean values for word amplitude, fundamental frequency, word duration, and spectral tilt for the quiet and 90 dB noise conditions. In general, these results replicated our previous findings using the Air Force vocabulary. For each of the three speakers in the hVd study, the presence of masking noise produced increases in word amplitude, fundamental frequency, and word duration. In addition, for each speaker, a decrease in spectral tilt occurred when utterances were produced in noise.

Insert Figures 5, 6, 7, and 8 about here

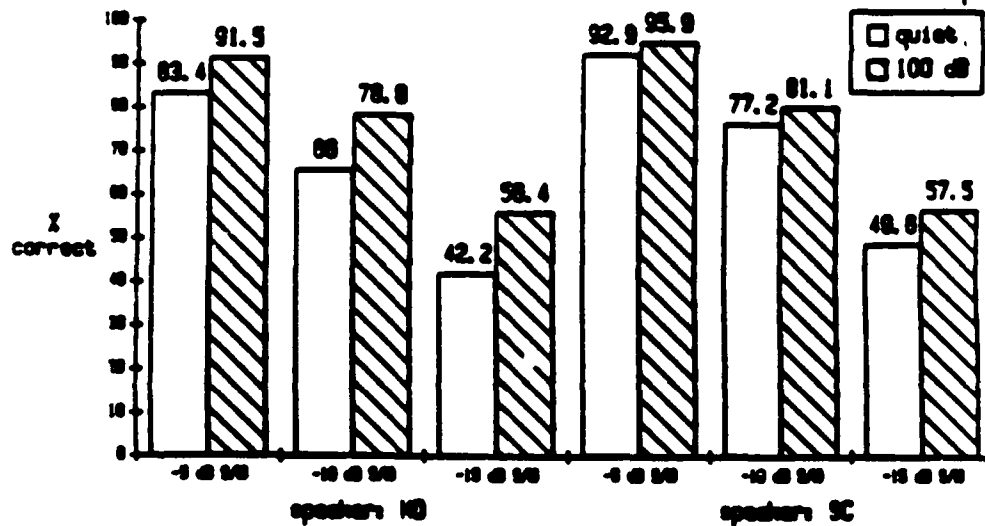
The consistent increase in fundamental frequency in the 90 dB noise condition of this study is particularly important because the results for F0 were not consistent across speakers in the earlier study using the Air Force vocabulary. In the initial study, only one of the two talkers showed a consistent increase in pitch as noise level increased. These new results suggest that increases in pitch are a common response to increased ambient noise in the environment.

As already mentioned, the increase in word duration found in the 90 dB noise condition of this study replicated our previous findings on word duration (Summers et al., 1988). We also examined durations of segments

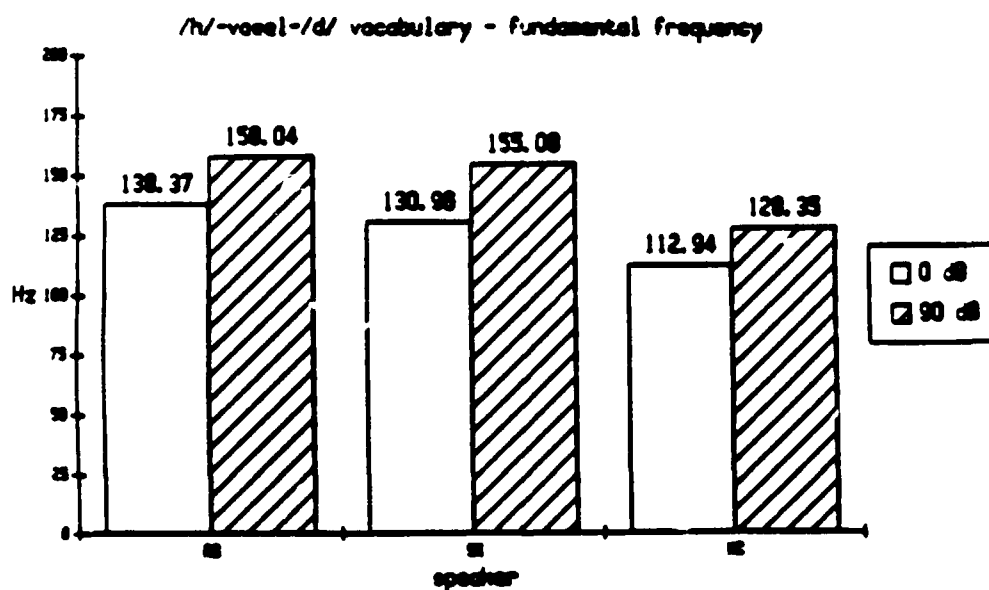
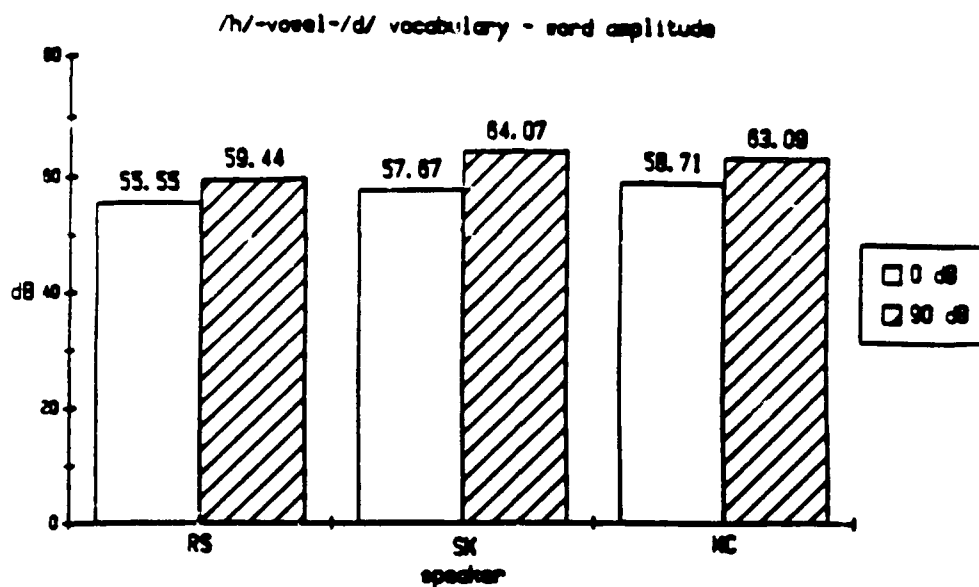
Expt 1: percentage of digits identified in noise



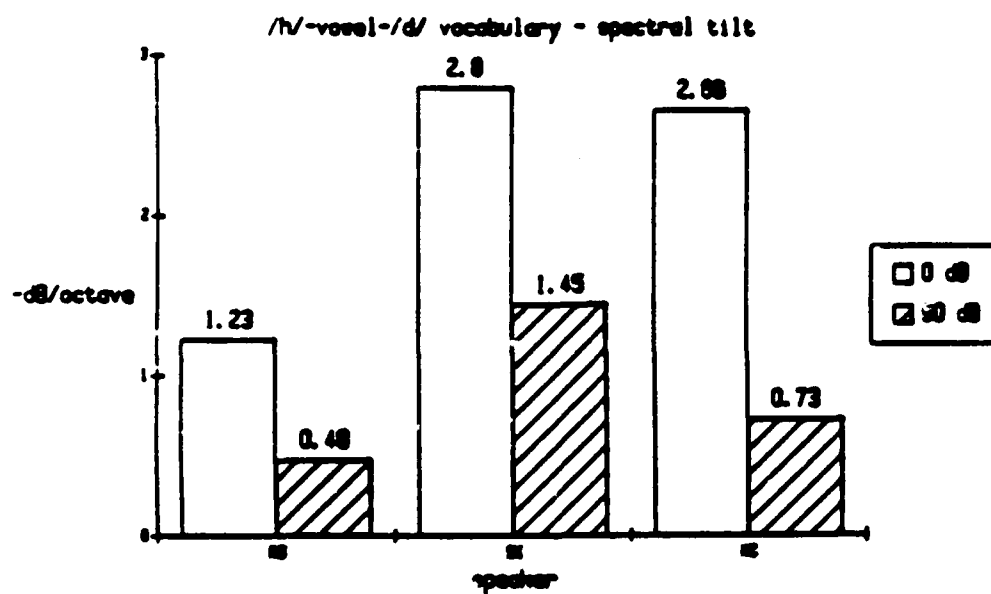
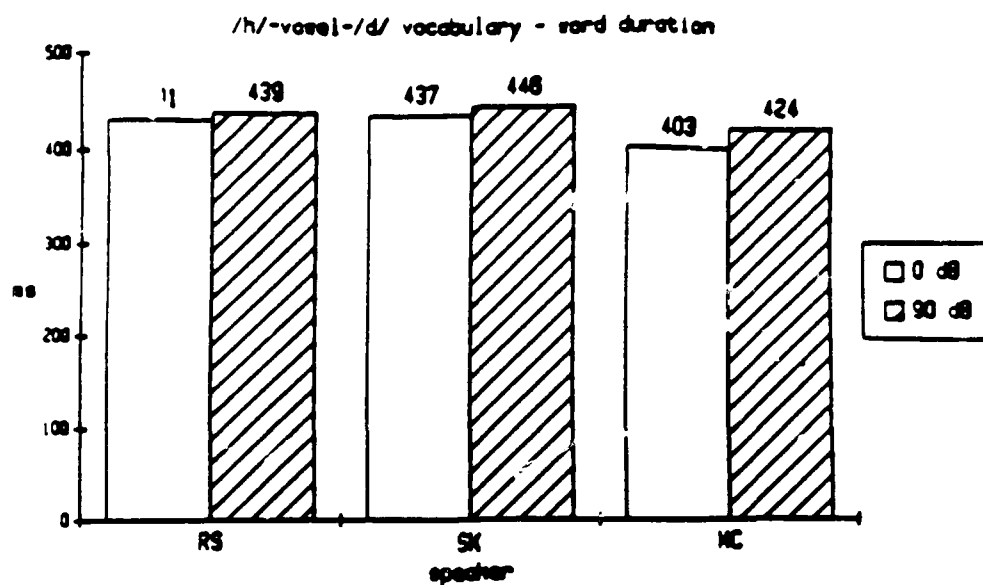
Expt 2: percentage of digits identified in noise



Figures 3 and 4. Intelligibility of digits produced in quiet and 90 dB of masking noise (Experiment 1) and digits produced in quiet and 100 dB of masking noise (Experiment II). Performance is broken down by S/N ratio and talker.



Figures 5 and 6. Mean values for word amplitude and fundamental frequency for words produced in quiet and 90 dB of masking noise by talkers RS, SK, and MC (/h/-vowel-/d/ vocabulary).



Figures 7 and 8. Mean values for word duration and spectral tilt for words produced in quiet and 90 dB of masking noise by talkers RS, SK, and MC (/h/-vowel-/d/ vocabulary).

within the hVd utterances in order to determine how this word lengthening is accomplished. For each utterance, the durations of three acoustic segments were measured: initial /h/ frication, vowel, and final closure prior to release of the final /d/. The duration of the final /d/ burst was not measured because the offset of this burst was too gradual to allow reliable segmentation from waveform displays. Mean durations for initial /h/ frication, vowel, and /d/ closure are broken down by speaker and noise condition in Figure 9. As the figure shows, the increase in total word duration for utterances produced in noise is due to an increase in vowel duration. There was no consistent pattern of a noise-related increase in /h/ frication or final closure duration for any speaker. In fact, there was a tendency for /h/ frication duration to decrease in the presence of noise.

Insert Figure 9 about here

Another focus of the hVd study was an examination of the effect of noise on the F1 and F2 frequencies for all 10 English monophthongs. Our previous analyses of the Air Force digit vocabulary suggested a tendency for F1 to increase in the presence of noise. However, this pattern was only present for one of two speakers in the previous study. Moreover, the vocabulary used did not contain all possible English monophthongs. We were therefore interested in examining the relationship between masking noise and F1 frequency in the productions of several additional speakers and in a more complete selection of vowels. Mean F1 and F2 frequencies for each of the ten hVd utterances are plotted separately for each speaker in Figures 10, 11, and 12. As these figures show, there was a consistent increase in F1 frequency for utterances produced in noise. The increase in F1 was present for all 10 vowels produced by each speaker. The figures also demonstrate that F2 frequencies were less affected by the presence of noise. For one of the three speakers (speaker SK), there was a tendency for F2 frequencies to be slightly lower in the presence of noise. This change in F2 was not observed for the other speakers.

Insert Figures 10, 11, and 12 about here

Talking in Noise: Consonant-vowel-/d/ Vocabulary

The results of the hVd study provided additional data on speech production in noise for a fairly simple vocabulary. In several further analyses, we examined speech produced in noise using a slightly more complex vocabulary. In these experiments, the initial consonant was varied along with the vowel in CVd syllables. The six stop consonants /b,d,g,p,t,k/ were used as initial consonants. These were paired with the vowels /i,a,ae,u/ in a consonant-vowel-/d/ context. Each of the six stop consonants was paired with each vowel for a total of 24 different utterances. This vocabulary allowed us to examine the influence of noise on the acoustic characteristics of voiced and voiceless consonants produced at various places of articulation. The consonant-vowel-/d/ vocabulary allowed us to test the reliability of our previous findings using a more complex stimulus set. This vocabulary also allows us to examine the influence of noise on initial formant transition

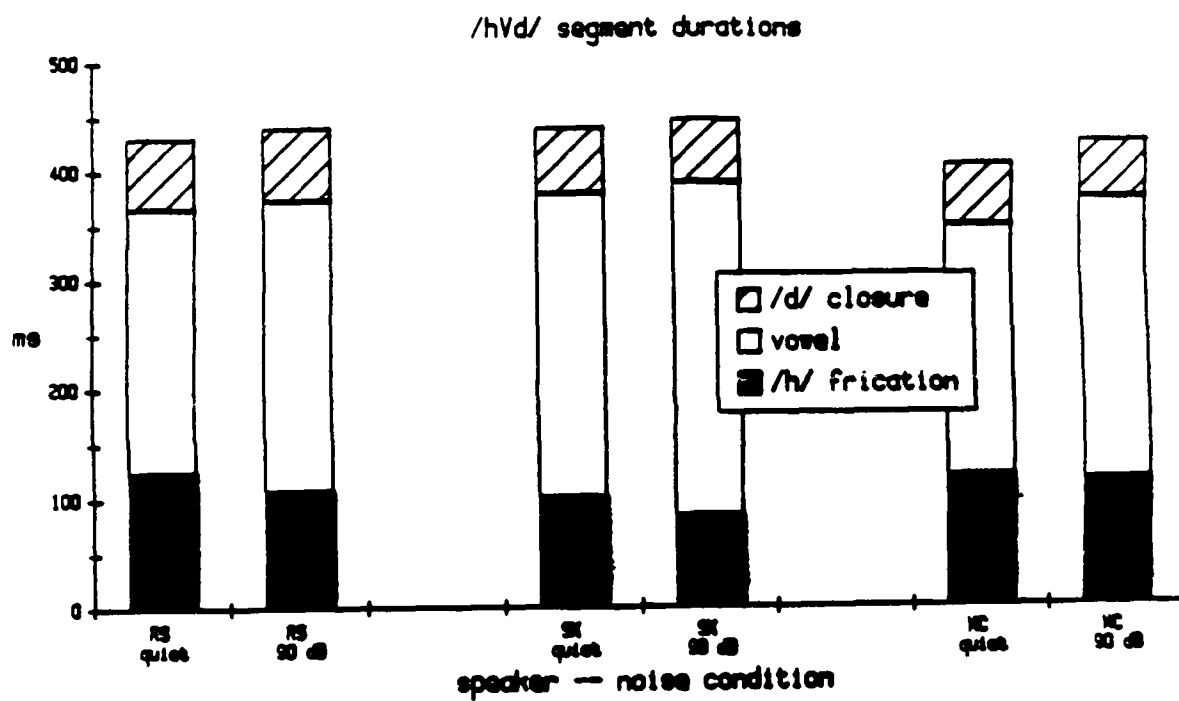


Figure 9. Mean durations for initial /h/ frication, vowel and /d/ closure for words produced in quiet and 90 dB of masking noise by talkers RS, SK, and MC (/h/-vowel-/d/ vocabulary).

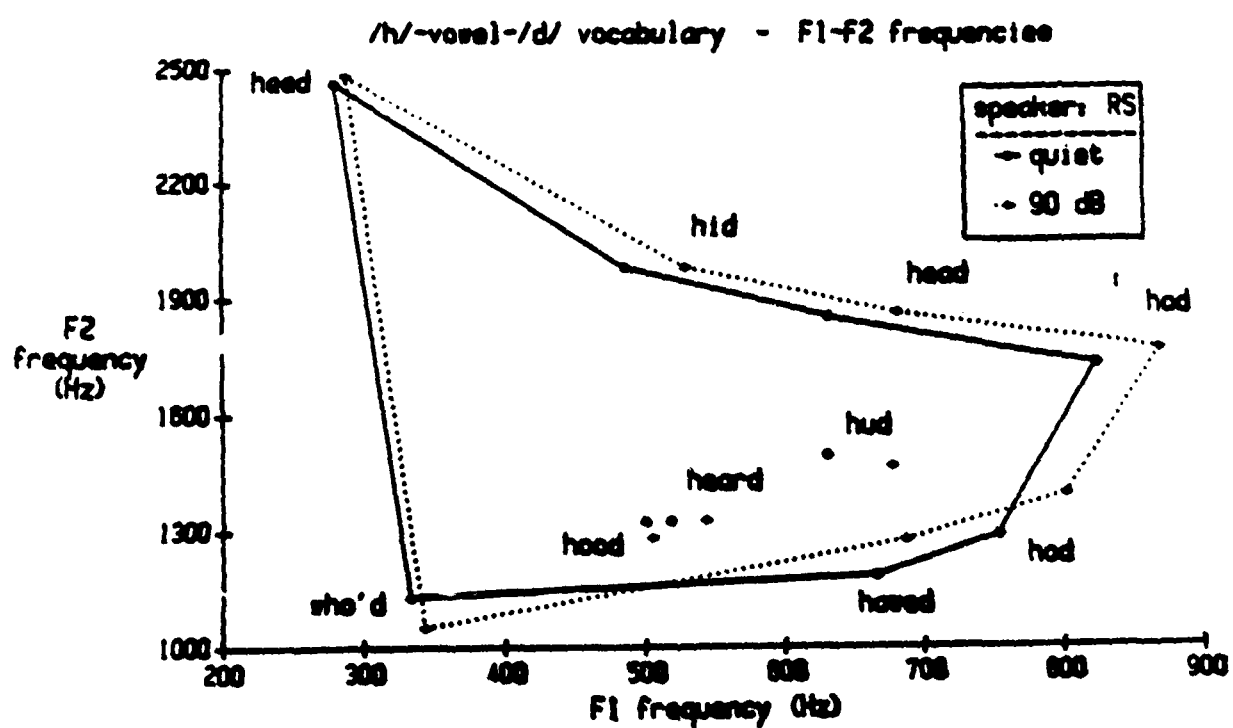


Figure 10. Mean F1 and F2 frequencies for each of the ten /h/-vowel-/d/ utterances produced in quiet and 90 dB by speaker RS.

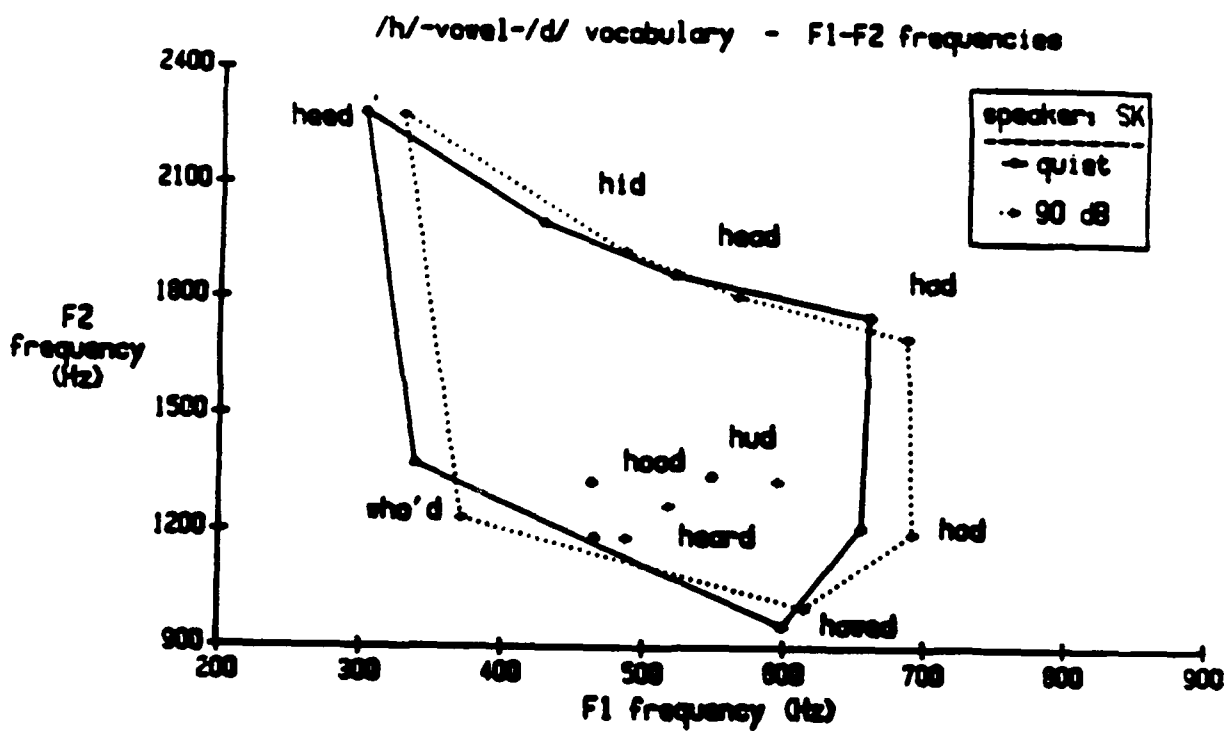


Figure 11. Mean F1 and F2 frequencies for each of the ten /h/-vowel-/d/ utterances produced in quiet and 90 dB by speaker SK.

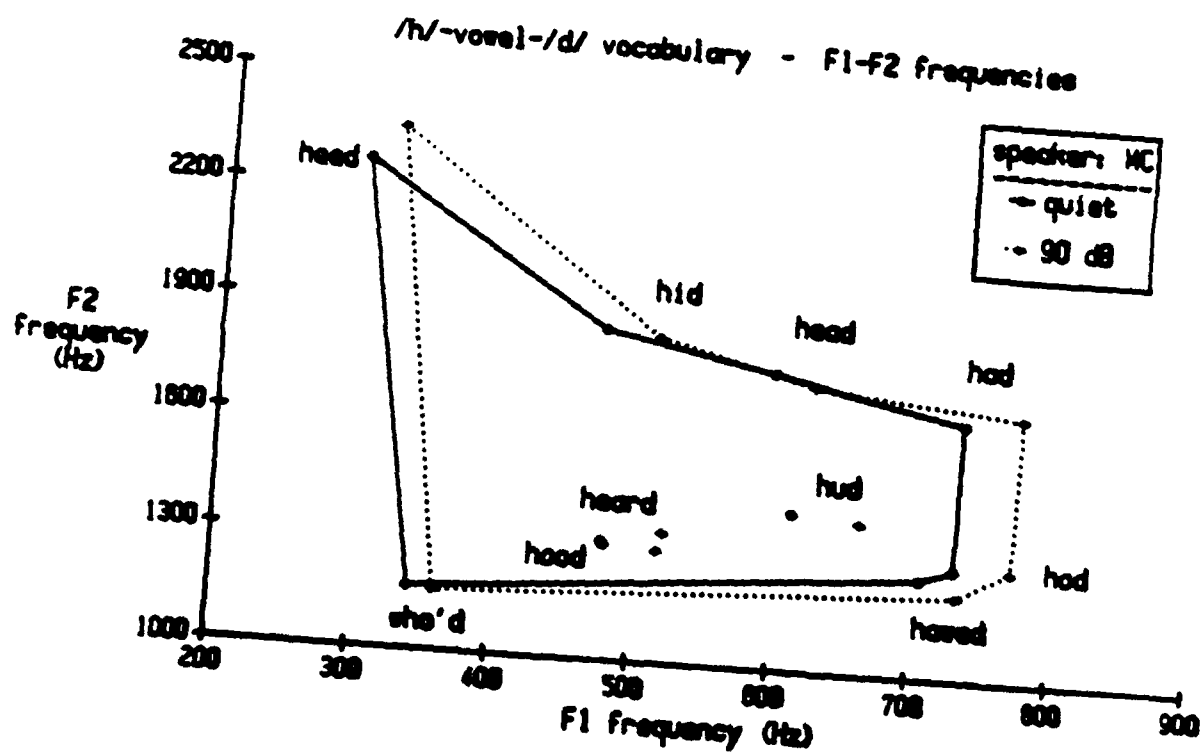


Figure 12. Mean F1 and F2 frequencies for each of the ten /h/-vowel-/d/ utterances produced in quiet and 90 dB by speaker MC.

regions associated with stop consonants in these environments.

We have analyzed the productions of the 24 consonant-vowel-/d/ utterances by two talkers who reproduced these utterances in the quiet and in 90 dB of masking noise. Our analyses of these utterances appear to replicate our previous findings that showed changes in amplitude, fundamental frequency, and spectral tilt. The results of these analyses of the consonant-vowel-/d/ utterances are shown in Figures 13, 14, and 15. Each speaker demonstrated an increase in amplitude and a decrease in spectral tilt for utterances produced in 90 dB of noise. These results replicated our previous findings. However, only one of the two speakers showed a consistent change in fundamental frequency across noise conditions. For this speaker, there was an increase in fundamental frequency for utterances produced in 90 dB of noise. When considered in combination with the F0 results from the previous studies, these results suggest that noise-related increases in F0 will only be observed for some speakers. Other speakers apparently make other adjustments in their productions while maintaining a fairly constant F0. In none of our studies, however, were there cases of F0 decreases accompanying increases in environmental noise.

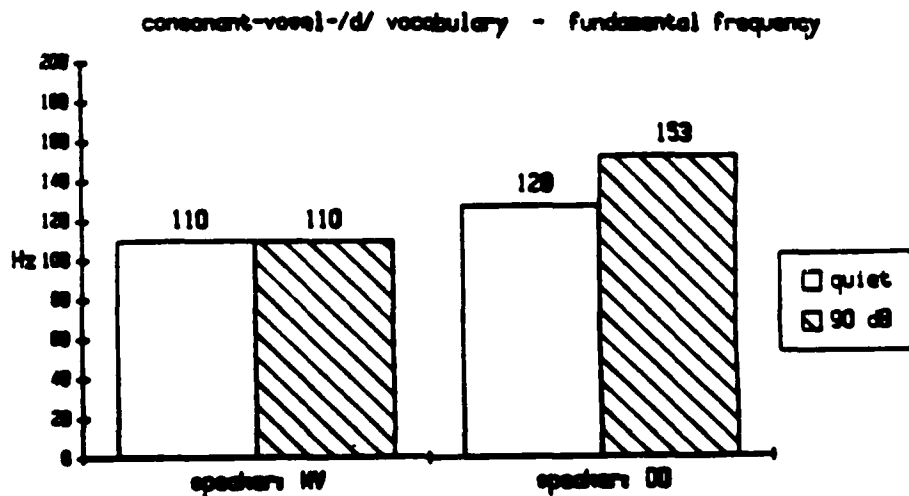
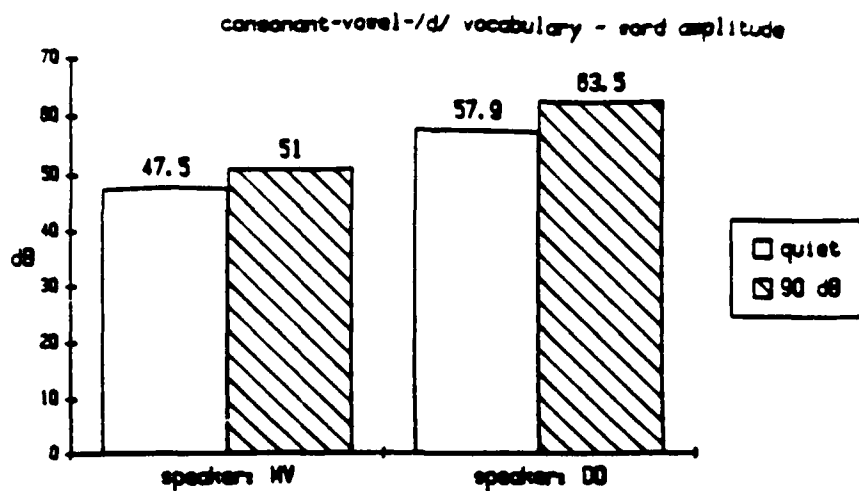
Insert Figures 13, 14, and 15 about here

The durational results for the consonant-vowel-/d/ vocabulary were of particular interest given our previous results with the /h/-vowel-/d/ vocabulary. In that study, increases in utterance duration found in the 90 dB noise condition were entirely due to increases in vowel duration. We were interested in determining whether this pattern was replicated in the consonant-vowel-/d/ vocabulary, where the vowel was not the only portion of the utterance that varied from trial-to-trial.

Prior to presenting the durational results, there is an issue concerning consonant production which must be addressed. During speech production, utterances containing initial /b/, /d/, or /g/ may be prevoiced so that the vocal cords begin to vibrate prior to consonantal release. In this case, vowel onset occurs at the time of release and there is no initial burst prior to vowel onset. One of the two speakers (DD) had a tendency to prevoice many of his utterances. Figure 16 displays the number of cases of prevoicing for utterances produced by speaker DD in the quiet and 90 dB noise condition. As the figure shows, this speaker was more likely to use prevoicing in the 90 dB noise condition than the quiet condition. Note that this increase in prevoicing can be seen as a method of increasing vowel duration while reducing initial consonant duration.

Insert Figure 16 about here

We now turn to the durational results. Each consonant-vowel-/d/ utterance was segmented into three components: an initial burst, a vowel segment, and a closure segment preceding the release of the final /d/. Mean durations for each of these acoustic segments are shown for each speaker and noise condition in Figure 17. The mean values shown in this figure do not



Figures 13 and 14. Mean values for word amplitude and fundamental frequency for words produced in quiet and 90 dB of masking noise by talkers MV and DD (consonant-vowel-/d/ vocabulary).

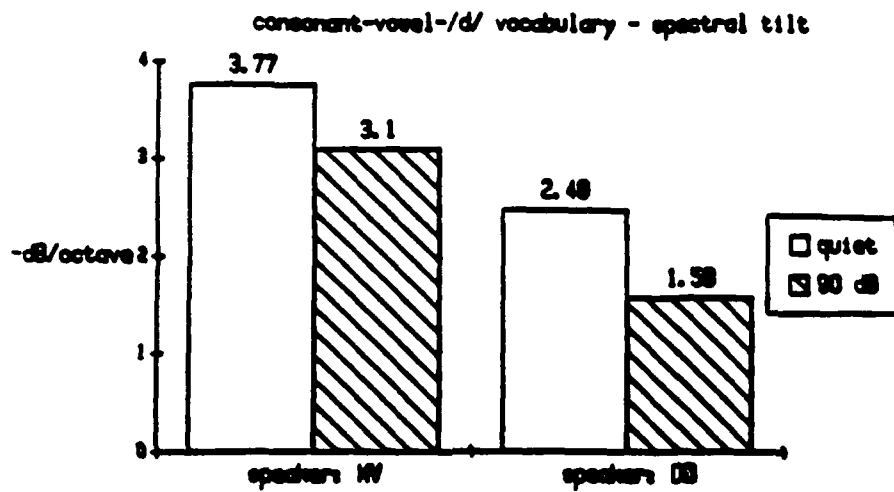


Figure 15. Mean values for spectral tilt for words produced in quiet and 90 dB of masking noise by talkers MV and DD (consonant-vowel-/d/ vocabulary).

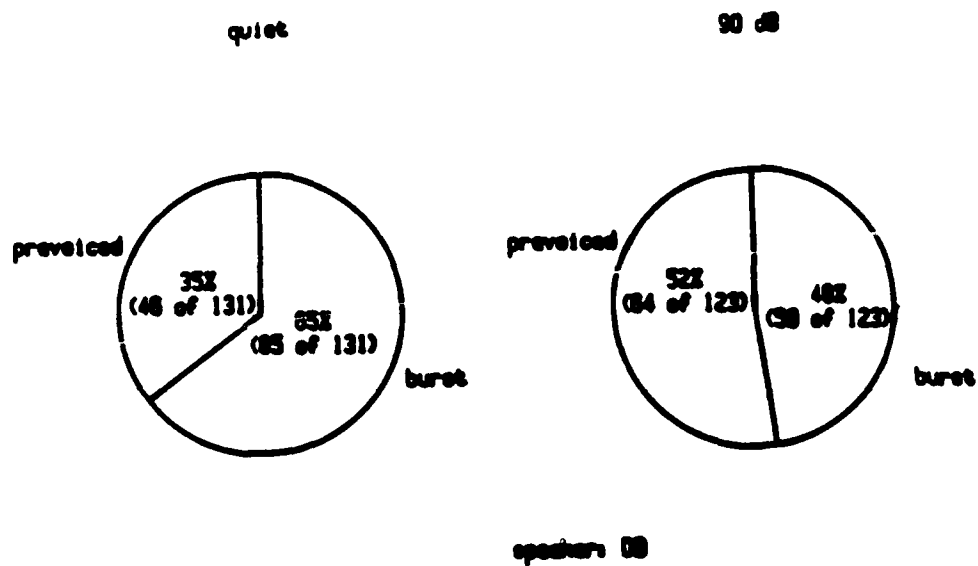


Figure 16. Number of cases of prevoicing for words containing initial /b/, /d/, or /g/, produced in quiet and 90 dB of masking noise by speaker DD (consonant-vowel-/d/ vocabulary).

include values from utterances containing prevoicing. These durational results are similar to those from the /h/-vowel-/d/ study. For each speaker, the increase in utterance duration seen in the 90 dB noise condition appears to be due to an increase in vowel duration. Initial burst durations and final closure durations did not increase in the 90 dB noise condition. In fact, there was a tendency for initial burst durations to decrease in the presence of noise. Thus, it appears that utterance lengthening due to masking noise is mainly confined to the vowel portions of these utterances.

Insert Figure 17 about here

The consonant-vowel-/d/ utterances were also examined in terms of F1 and F2 frequencies. Mean F1 and F2 frequencies collapsed across initial consonants are shown for each speaker in Figures 18 and 19. In each case, the F1 increases in the presence of noise. This pattern was present for all vowels produced by each speaker. However, the results for F2 were not consistent across speakers. For MV, F2 frequencies tended to decrease in the presence of noise. Speaker DD showed exactly the opposite pattern with increased F2 frequencies for utterances produced in noise. In summary, the F1 and F2 results replicated our previous findings of F1 increases accompanying increased noise in the environment. As in our previous studies, F2 frequencies did not show a consistent pattern of change across noise conditions.

Insert Figures 18 and 19 about here

It should be mentioned here that some objections have been raised to our method of LPC formant estimation (Fitch, 1989). In a letter to the editor of JASA, Fitch claimed that LPC analysis is biased towards picking the nearest harmonic of the fundamental frequency for its estimate of F1. We refuted this argument by demonstrating an inconsistent relationship on a token by token basis between the value of F1 and the value of F0 derived from LPC analysis (see Summers et al., 1989). Fitch also argued that F1 estimates may be affected by spectral tilt. She stated that the amplitude increases related to spectral tilt may influence the LPC model to pick values of F1 which are inaccurate. This argument was refuted by an analysis of covariance which demonstrated that the increase in F1 is not completely accounted for by the variability in spectral tilt. Reanalysis of our earlier data using different preemphasis weights and manipulated tilt measures demonstrated that spectral differences between speech produced in a quiet environment versus speech produced in noise still existed, contrary to Fitch's arguments (see Summers et al., 1989).

In summary, the results of our work on the effects of noise on speech production indicate that a wide variety of changes occur at the acoustic-phonetic level. The first and most obvious is an increase in amplitude when talkers speak in a noisy environment. There is also an increase in duration and a decrease in spectral tilt. Utterances produced in noise become louder, longer and more monotone. F0 tends to rise, as does F1. These two effects were shown to be independent of one another. In contrast, F2 remains

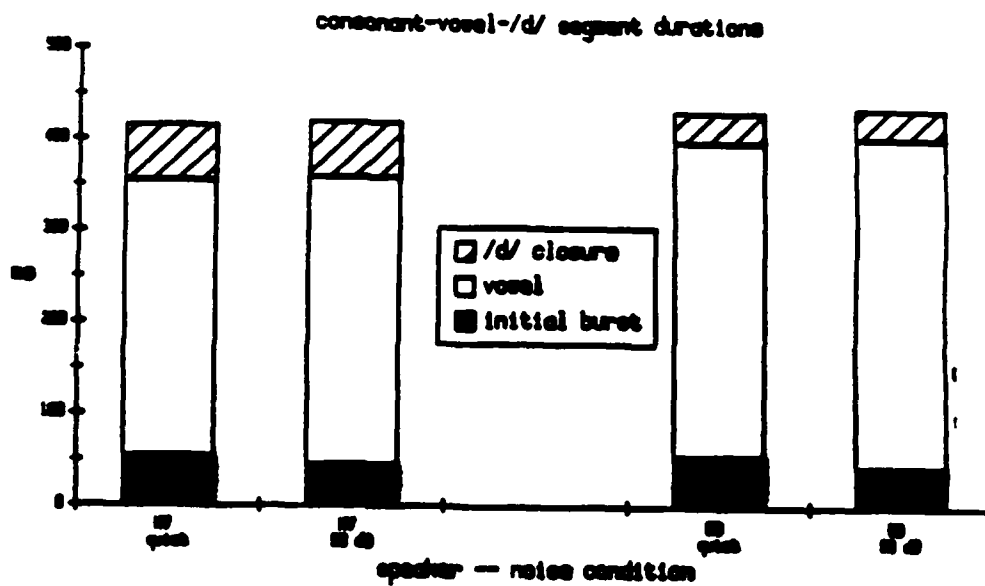
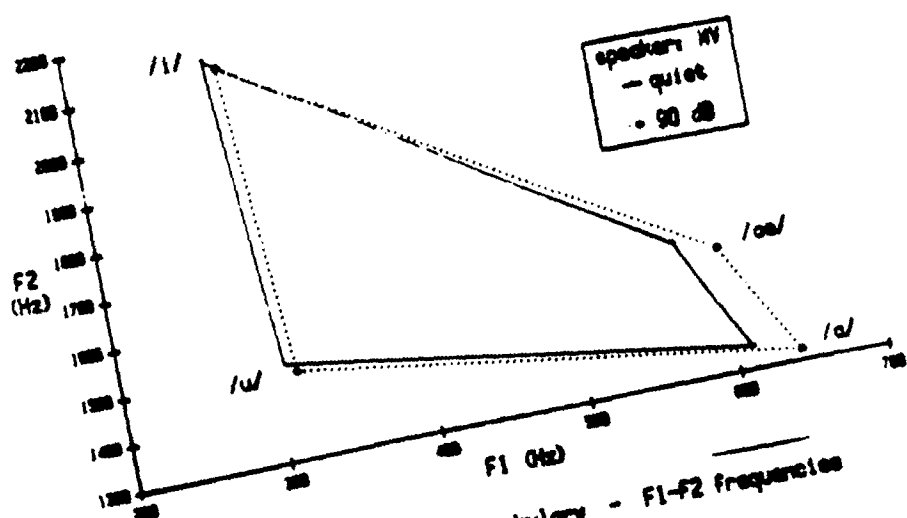
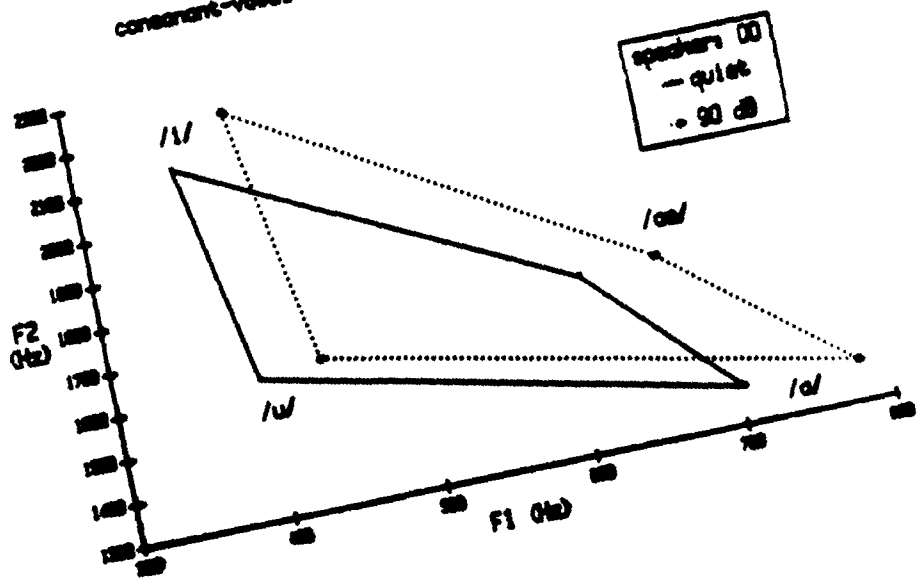


Figure 17. Mean durations for initial burst, vowel, and /d/ closure for words produced in quiet and 90 dB of masking noise by talkers MV and DD (consonant-vowel-/d/ vocabulary).

consonant-vowel-/d/ vocabulary - F1-F2 frequencies



consonant-vowel-/d/ vocabulary - F1-F2 frequencies



Figures 18 and 19. Mean F1 and F2 frequencies collapsed across initial consonants for words produced in quiet and 90 dB of masking noise by talkers MV and DD (consonant-vowel-/d/ vocabulary).

relatively unaffected. Taken together, these results replicate and extend in several important ways the previous research on the Lombard effect (Lane et al., 1970; Lane and Tranel, 1971; Hanley and Steer, 1949; Draegert, 1951; Pisoni et al., 1985; Summers et al., 1988).

Talking Under Acceleration: The AAMRL Tapes

The speech recorded at AAMRL from two talkers speaking under acceleration was also analyzed during the contract period. The 15-word Air Force vocabulary was produced by two talkers under the following conditions: 2-g, 3-g, 4-g, 5-g, and 6-g's. Each talker produced five tokens of each word in each condition. The speech of one talker was digitized and edited into individual stimulus files for acoustic analysis and subsequent use in perceptual experiments. Some words only had four tokens whereas other words had seven tokens. In the 6-g condition, only three tokens for each word were recorded.

We carried out acoustical analyses of the speech from one talker. Mean values for word amplitude, fundamental frequency, word duration, and spectral tilt were examined for each of the five acceleration conditions. Our results showed little influence of acceleration on word duration. Mean durations were similar in the 2-g, 3-g, 5-g and 6-g conditions and slightly lower in the 4-g condition. Acceleration had a more pronounced effect on the other three measures. Our analyses showed an increase in fundamental frequency of approximately 10 Hz in the 4-g, 5-g, and 6-g conditions in comparison to the 2-g and 3-g conditions. Mean pitch values were similar in the 2- and 3-g conditions and in the 4-, 5-, and 6-g conditions.

Increases in word amplitude were also observed. This pattern held across the 2-g, 3-g, 4-g, and 5-g conditions with a small decrease in amplitude from the 5-g to the 6-g condition. Mean amplitude in the 6-g condition was greater than in the 2-g, 3-g, or 4-g conditions.

Finally, changes in acceleration appear to have had some influence on spectral tilt. We found greater tilt in the 2-g, 3-g, and 4-g conditions than in the 5-g and 6-g conditions. Increases in acceleration produce shifts in the relative distribution of spectral energy so that higher frequencies increase in amplitude relative to lower frequencies.

Analyses of variance on these data showed a significant main effect of acceleration for each of the dependent variables examined: word amplitude, F0, word duration, and spectral tilt, ($p < .05$ for each dependent variable).

Perceptual Experiments on Acceleration

We also completed two perceptual experiments that examined whether the acoustic-phonetic changes in speech produced under differing conditions of acceleration could be readily identified by human observers in controlled listening environments. These experiments used the speech produced in the 2-g and 5-g acceleration conditions described in the previous section. For each of the 15 stimulus words, three tokens from the 2-g condition and three tokens from the 5-g condition were used. Each token of each stimulus word produced at the 2-g acceleration level was paired with each token of each word produced at the 5-g level in a paired-comparison testing format.

Subjects were instructed to listen to each pair of words and to identify the pair member produced in the 5-g environment. They were told that the higher g force present in the 5-g environment made this environment more stressful than the 2-g environment and that they should identify the pair member produced in the "high stress" environment. In the first experiment, the stimuli were presented to the subjects exactly as each had been spoken by the talker under the two acceleration levels; that is, the tokens were not normalized for amplitude. Thus, amplitude differences were present along with all other cues to aid subject's judgments. Across all subjects, the 5-g stimuli were labelled "high stress" in 79.7% of the trials. This result is significantly better performance than the 50% level of accuracy that would be expected if subjects were simply guessing, ($t(8) = 5.18, p=.0008$).

To determine whether the results of this experiment were due to amplitude differences between the 2-g and 5-g stimuli, a second perceptual experiment was carried out in which amplitude differences between utterances were removed. Stimulus materials for the second experiment were identical to those used in the first experiment with one modification: all stimuli used in this experiment were equated in terms of overall RMS amplitude. Identification performance in the second experiment demonstrated again that the 5-g stimuli could be reliably labelled "high stress" on 77.9% of the trials. This is significantly better than expected by chance, ($t(8) = 14.83, p<.0001$).

In comparing the results of the two perceptual experiments, the 5-g stimuli were identified as "high stress" slightly more often in Experiment 1 than in Experiment 2 (79.7% versus 77.9% of the time). Given that labelling performance was significantly better than chance in both experiments and that overall performance was only slightly better in Experiment 1 than in Experiment 2, it seems clear that amplitude differences are not the only source of information available to subjects in identifying which items were produced under "high stress". In other words, reliable changes in the acoustic-phonetic properties of speech appear to be present in these speech waveforms in addition to any changes in vocal level.

Project II. Effects of Cognitive Load on Speech Production: The JEX Task

Attentional and cognitive demands placed on pilots, flight controllers, and other operators involved in information-intensive jobs may influence the acoustic characteristics of their speech in demanding situations. Very little research has explored whether consistent changes can be identified in the characteristics of utterances produced in demanding or "high-workload" environments. Knowledge of these changes in speech production could have several important implications. First, if properties of speech could be identified which correlate with the level of workload an operator is experiencing, this information could be used in training and selecting operators or in testing environments for their human-factors acceptability. This information could also be important in the design of speech-recognition devices that are used in high-workload settings. Robust recognition systems must be able to tolerate changes in acoustic characteristics of speech that occur as a result of variability in workload. This project was designed to explore whether consistent changes in speech could be identified which were the result of changes in the attentional and cognitive demands of the environment.

Previous research involving workload tasks has generally assumed that workload increases are associated with increased psychological stress (e.g., Hecker, Stevens, von Bismark and Williams, 1968; Tolkmitt and Scherer, 1986). The results of these studies have often been equated with studies in which psychological stress is manipulated through exposure to aversive stimulation, instructions requiring subjects to lie to the experimenter (or an accomplice) or some other means of increasing emotional stress (Scherer, 1979). In addition, the majority of previous studies concerned with these issues have focused on fundamental frequency (F0) characteristics as an indicator of workload or stress (see, for example, Williams and Stevens, 1969; Kuroda, Fujiwara, Okamura and Utsuki, 1976; Tolkmitt and Scherer, 1986). Few studies have examined a larger array of acoustic characteristics in an effort to produce a more complete description of the effects of increased workload on the speech signal (but see Hansen, 1988). The present investigation examined how changes in cognitive workload affect several acoustic characteristics of the speech signal and whether changes that can be associated with increased workload are similar to changes that have previously been ascribed to increased emotional stress.

Cognitive workload was manipulated by requiring speakers to perform an attention-demanding secondary task while reading lists of test words from a CRT display. The task chosen was a compensatory tracking task which was first described in JEX, McDonnell and Phatak (1966). The tracking task will be referred to as the "JEX" task hereafter. The task involved manipulating a joystick in order to keep a pointer centered between two boundaries on a computer screen (see Figure 20). The program deflected the pointer away from the center position and the subject was required to continuously compensate for the movement of the pointer by manipulating the joystick in order to keep it from crashing into one of the boundaries. Phrases that the subjects were required to produce were presented visually on the computer screen while the subjects continued to perform the JEX task.

Insert Figure 20 about here

The influence of cognitive workload on various acoustic characteristics of the test utterances is described below. In each case, an analysis of variance was used to determine whether workload had a significant effect on a given acoustic measure. Separate analyses were carried out for each speaker. The analyses used phrase and workload condition (JEX versus control) as independent variables. The presentation of the results will focus on the effect of workload on the various acoustic measures.

Amplitude

The upper panel of Figure 21 shows vocal amplitudes averaged across entire phrases for utterances from the JEX and control conditions. The data are plotted separately by speaker. The lower panel of the figure shows amplitudes at the segmental level. Amplitudes of the /h/ frication, vowel, and /d/ closure portion of each hVd utterance are shown for each workload condition. An asterisk above a pair of bars indicates a significant difference ($p < .05$) between values in the JEX and control conditions for a particular speaker.

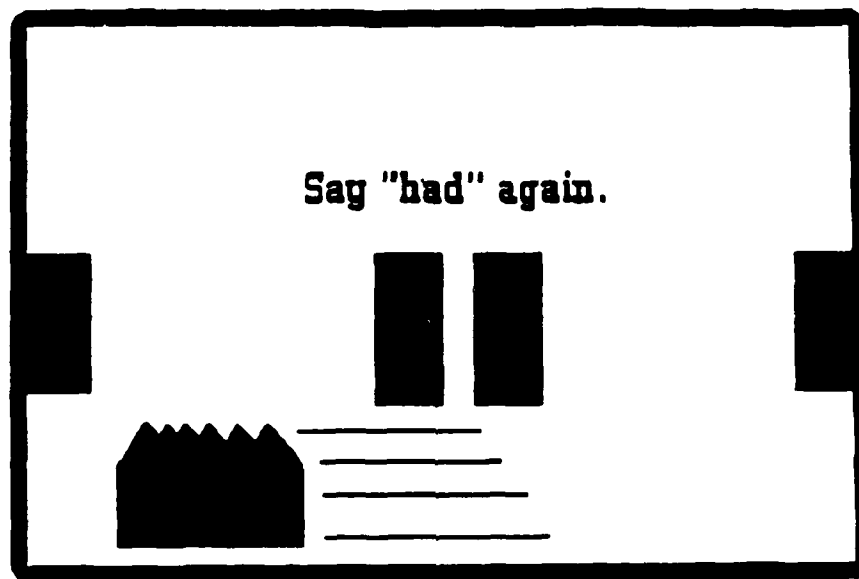


Figure 20. Illustration of the Jex compensatory tracking procedure. The subject's task is to keep the moving pointer located at the bottom of the display from crashing into the sides of the display. During speech-collection intervals, sentences were presented in the top portion of the display as shown.

Insert Figure 21 about here

As the figure shows, there was a tendency for amplitude to increase in the JEX condition. The pattern is very consistent for speakers SL, ME, and TG, who showed significantly higher amplitudes in the JEX condition for the entire phrase and for the separate /h/, vowel, and /d/ closure segments. Speaker MC showed significantly higher amplitudes in the JEX condition for the vowel and /d/ closure segments. Speaker EG did not show as clear a pattern of amplitude increases under workload as the other speakers. For this speaker, /h/ frication amplitude was significantly higher in the control condition than in the JEX condition. However, this significant main effect was mediated by a significant phrase X workload condition interaction. For EG, /h/ frication amplitude was higher in the control condition in 7 of the 10 vowel contexts. Of all of the analyses carried out in this study, this was the only case of a significant phrase X workload interaction.

Amplitude Variability

Along with an increase in mean amplitude, amplitude variability from one utterance to the next also tended to increase in the workload condition. Figure 22 shows standard deviations of phrase amplitude across utterances for each condition. For the entire phrase, four of the five subjects showed an increase in amplitude variability in the JEX condition. For three of these subjects, the increase in variability was statistically significant. One subject showed the opposite pattern with significantly less amplitude variability in the JEX condition. As the lower panel of the figure shows, the pattern just described for the entire phrase was also true for amplitude variability of vowels and /d/ closure segments in the h-vowel-d context. In each case, four of the five subjects showed greater amplitude variability when performing the workload task. The effect was statistically significant for three of the four speakers in the case of vowels but was only significant for one speaker in the case of /d/ closure.

Insert Figure 22 about here

Amplitude increases are often correlated with changes in spectral tilt. That is, high amplitude utterances generally show flatter spectra with relatively more high frequency energy than is found in low amplitude utterances. We examined the long-term spectra of the hVd vowels in each condition to determine if the amplitude increases found in the JEX task were correlated with decreased spectral tilt. Figure 23 shows the difference in energy between JEX and control condition vowels across 40-Hz linear frequency bands.

Insert Figure 23 about here

A positive slope in these figures indicates that the difference in energy between JEX and control condition vowels increased with frequency. This pattern is present for speakers MC, SL, ME, and EG. Thus, as in the amplitude

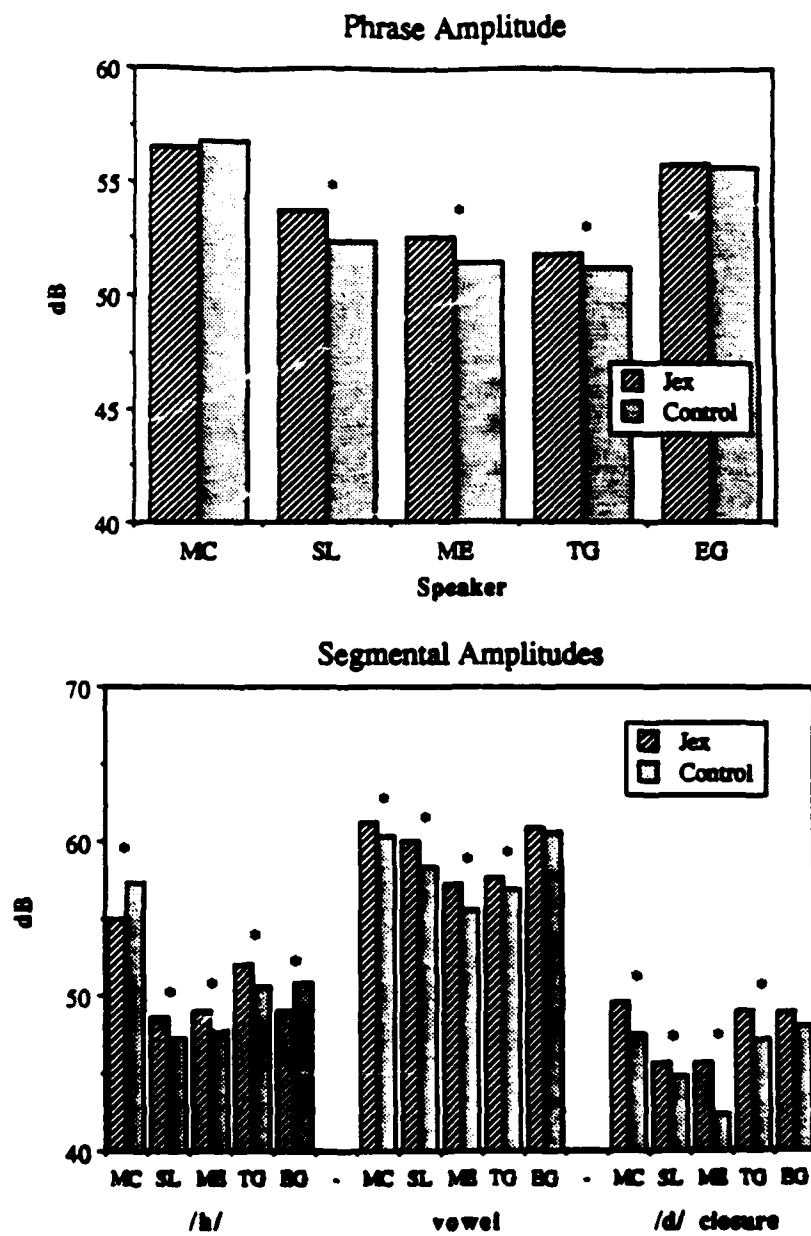


Figure 21. Phrase amplitude (upper panel) and segmental amplitudes for "Say hVd again" utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

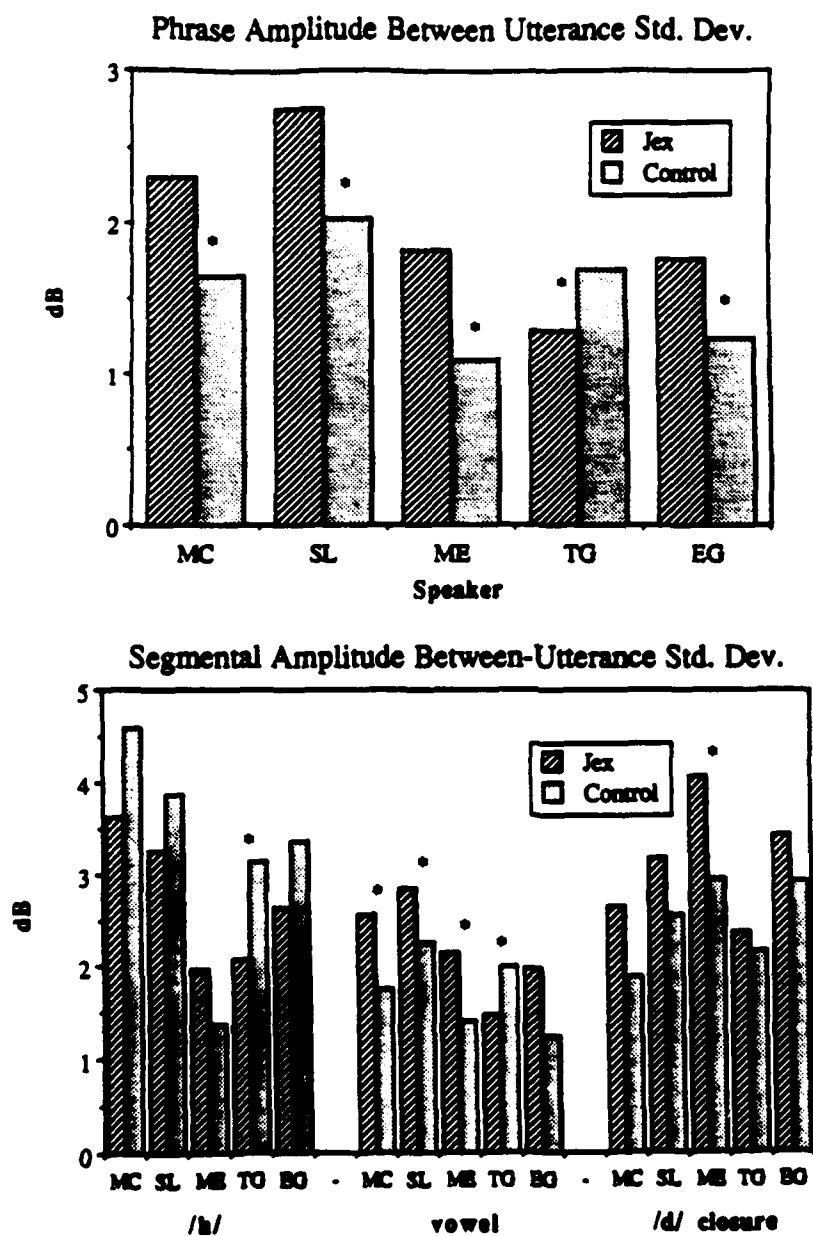


Figure 22. Between utterance standard deviations for phrase amplitudes (upper panel) and segmental amplitudes. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

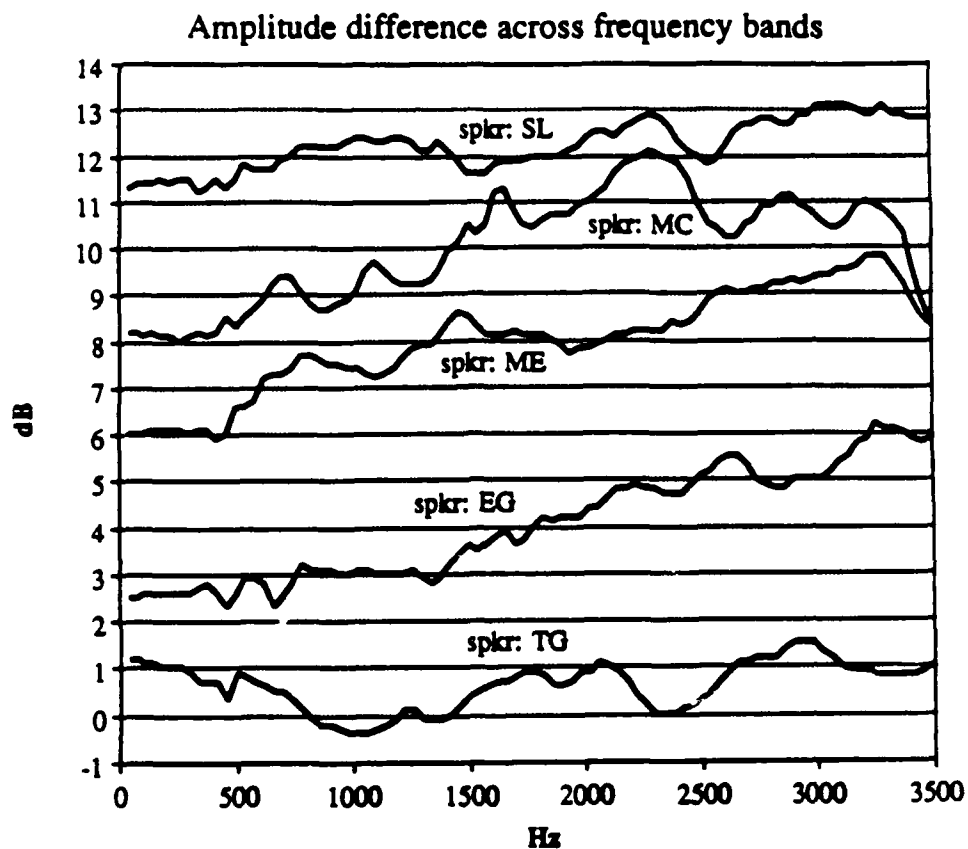


Figure 23. Mean differences in energy between utterances produced in the JEX and control conditions across frequency bands. Values are collapsed across utterances and presented separately for each speaker. For clarity, the traces for speakers EG, ME, MC and SL have elevated by 2.5, 5, 7.5 and 10 dB respectively.

data, four of the five subjects showed a consistent pattern. However, the four subjects who showed changes in spectral tilt across conditions are not the same as the four who showed amplitude differences. Subject EG, the subject who did not show significant amplitude differences across workload conditions, provided one of the clearest cases of changes in spectral tilt. Overall, the data suggest that the workload task produced effects on spectral tilt that were not always correlated with changes in overall amplitude.

Fundamental Frequency

We also analyzed fundamental frequency for each phrase and for the hVd vowel segments in each condition. Figure 24 shows mean F0 values for the phrase and hVd vowels in each condition for each talker. Two speakers (MC and SL) showed a significant increase in F0 for the entire phrase and for the hVd vowel in the JEX task. The pattern of F0 increase under workload was not replicated for either the phrase or hVd vowel in the data from the other three subjects.

Insert Figure 24 about here

Fundamental Frequency Variability

A different property of the F0 data did, however, show a fairly consistent pattern across workload conditions. Figure 25 shows the standard deviations of the frame-by-frame F0 values for each phrase and for each hVd vowel in each condition. As the figure shows, three subjects showed a significant reduction in F0 variability when performing the JEX task. A fourth subject showed the same general pattern, although the difference was not significant. The pattern was not observed in the vowel data. Given that F0 variability decreased for the entire phrase but not the vowel, the pattern is more likely to be due to a flattening of the overall F0 contour rather than a decrease in period-to-period F0 variability or "vocal jitter." In other words, the whole phrase is apparently produced using a monotone pitch when the subject is under workload.

Insert Figure 25 about here

Vocal Jitter Measures

To rule out the possibility that the decrease in F0 variability in the JEX condition was due to a decrease in "vocal jitter," acoustic analysis was performed on each /a/ and each /uh/ produced in the JEX task. The jitter in these utterances was compared to the jitter in the subjects' utterances produced in the control condition. Statistical analysis revealed no significant difference in jitter between stress conditions for any of talkers. Based on this analysis, we can tentatively rule out a decrease in pitch period to pitch period variation as the cause of the decrease in F0 variation.

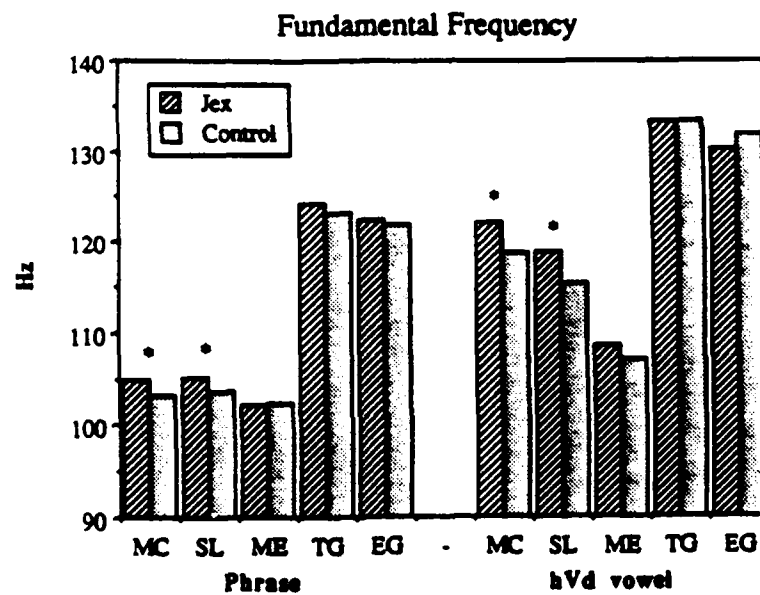


Figure 24. Mean fundamental frequency values for utterances produced in Jex and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

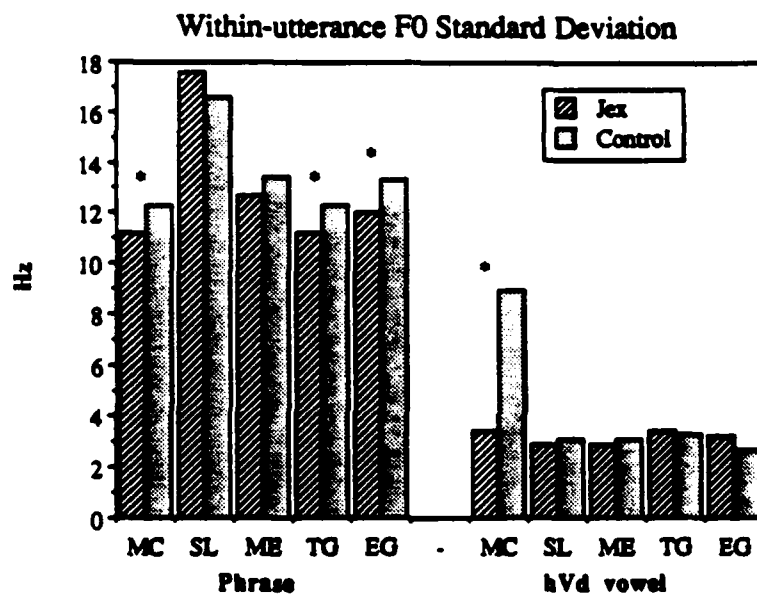


Figure 25. Mean within-utterance F0 standard deviations for utterances produce in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

Duration

Figure 26 shows the effect of cognitive workload on phrase durations and segmental durations. Four of the five speakers showed significantly shorter overall phrase durations while performing the JEX task. One speaker showed the opposite pattern with longer phrase durations under workload. This speaker showed the smallest change in phrase duration across conditions. Segmental durations also tended to be reduced while performing the JEX task. The four speakers who showed shorter phrase durations in the JEX condition also tended to show shorter /h/ frication durations and shorter /d/ closure durations. Vowel duration (in hVd words) was less consistently affected by the workload condition. The durational shortening observed for the entire phrase, the /h/ frication, and the /d/ closure, replicate results mentioned briefly by Hecker et al. (1968).

Insert Figure 26 about here

Given that the vowel in the hVd contexts was the only part of the phrase containing "new" information from trial-to-trial, speakers may have treated the production of this vowel as more important than the production of the surrounding context. This may explain why vowel duration was not consistently reduced in the JEX condition while other segmental durations were reduced in the remainder of the utterance.

Formant Frequencies

Workload did not have any consistent effects on the frequencies or bandwidths of the first three formants for any of the five speakers. Thus, it appears that workload had a greater influence on sub-laryngeal and laryngeal (source-related) functions and speech timing than it did on the supralaryngeal control of speech.

Summary and Conclusions

Very little previous research in the published literature has attempted to identify consistent changes that occur in the acoustic-phonetic properties of speech produced in severe environments. Research in this area may have important implications for human-to-human and human-to-machine speech communication in demanding environments such as cockpits and air traffic control towers. The present results show that increased cognitive workload produces a number of consistent effects on the acoustic-phonetic properties of speech. Utterances produced under cognitive workload show higher amplitudes and greater amplitude variability between utterances. Spectral tilt was reduced for vowels produced under workload but this change in tilt was not always correlated with a change in amplitude. F0 variability within an utterance was reduced under workload, suggesting that these utterances were produced with a flatter and perhaps less expressive F0 contour. Overall phrase durations and segmental durations were also reduced under workload, suggesting an overall increase in speaking rate as workload increased.

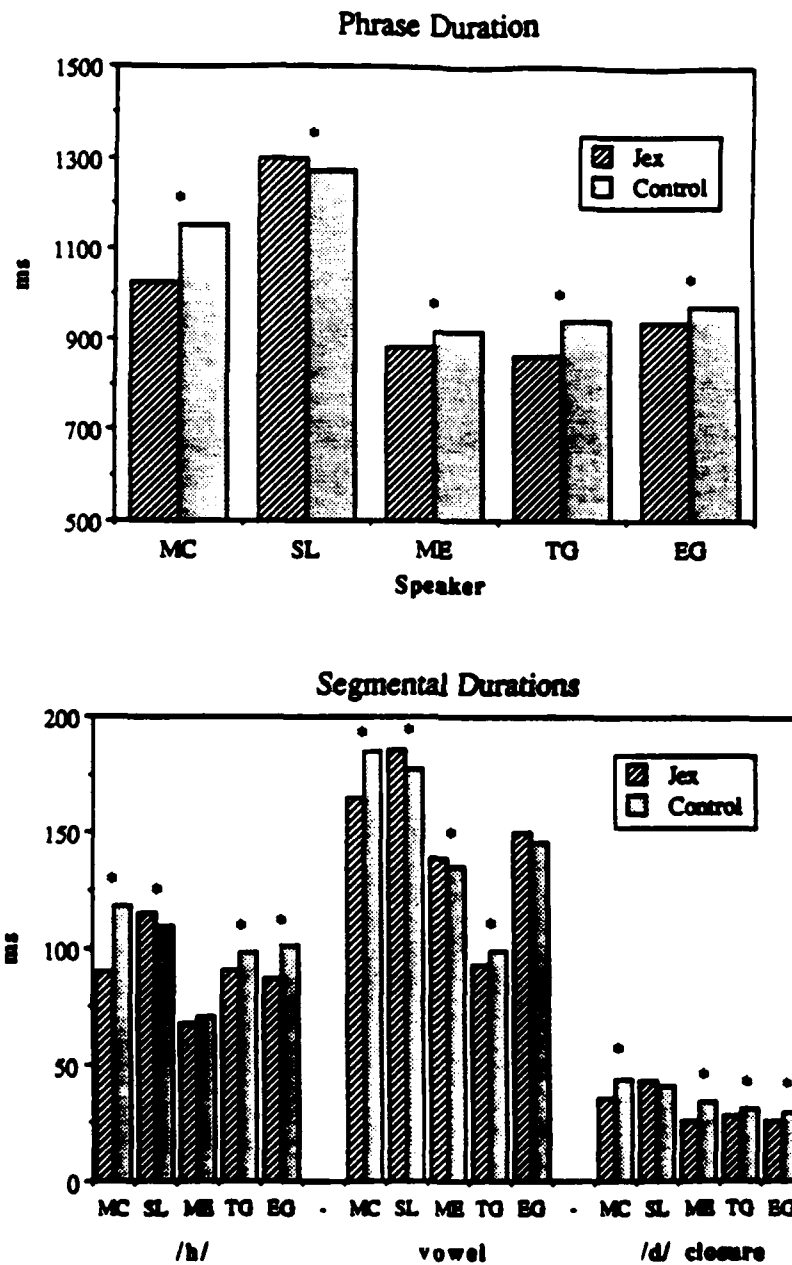


Figure 26. Mean phrase duration (upper panel) and mean segmental duration values for utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

The patterns reported here are tendencies that emerged from analyses of a small number of subjects. Some differences were not always present for each subject. We believe that these patterns may be more consistent in an actual "high workload" environment than could be seen in this investigation in which performing poorly on the workload task had only minor consequences (compare, for example, Williams and Stevens', (1968) analysis of F0 characteristics in tape-recordings of actual conversations between pilots and flight controllers to Hecker et al.'s (1968) analysis of F0 in a laboratory task designed to increase workload). Of the five speakers examined here, subject MC performed the JEX task at the highest level of difficulty and may have been the most highly motivated of the five subjects. It is interesting to note that, in general, MC showed the most consistent effects of workload on the acoustic-phonetic properties of his speech.

The absence of any effect of workload on formant frequencies in combination with the other findings suggests that the main effect of workload occurred at or below the level of the larynx. The changes in amplitude, F0 characteristics, and spectral tilt that we found in this study may be related to changes in the shape and variability of the glottal waveform (Hecker et al., 1968).

In summary, the results of this project demonstrate a number of reliable changes in the acoustic-phonetic properties of speech produced under increased cognitive workload. The findings add to a growing body of literature showing that talkers consistently modify their speech in response to both physical and mental changes in their immediate environments. These results have important implications for the use of speech recognition devices in severe environments in which the operator may be required to carry out several demanding activities at the same time.

Project III. Effects of Vocal Fatigue on Speech Production

The third project has been concerned with acoustic-phonetic changes that occur as a result of vocal fatigue. Very little research has been carried out to assess the time course of vocal fatigue and quantify the changes that occur over this period (Scherer et al., 1987; Sander and Ripich, 1983). Most of the research concerning vocal fatigue has focused on clinical populations (Yanagihara, 1967). These are talkers with preexisting vocal impairments who differ in their degree of subjectively rated hoarseness. The goal of previous research has been to find consistent subjective labels to describe the voice quality of the subjects. Several studies have attempted to quantify vocal fatigue in terms of a harmonics-to-noise ratio. As hoarseness increases, the ratio of the amplitude of the harmonics to the amplitude of the noise components should decrease. Noise components become more pervasive in the spectrum as hoarseness increases. This adds to the breathy, somewhat raspy quality of the vocally fatigued voice.

Our interest in vocal fatigue was to trace its time course and to look at the acoustic-phonetic correlates that occur as a result of vocal fatigue. We report the results of a pilot study in which we attempted to induce vocal fatigue in two subjects by having them speak at a loud level for an extended period of time. In order to address both segmental and prosodic effects of prolonged speech, target words included monosyllabic CVCs (e.g., pap, tat and

back) and hVds (e.g., hoed and who'd), as well as bisyllabic words which could be either a noun or verb, depending on which of the syllables was stressed (e.g., SUBject and subJECT).

One male and one female subject, both members of the laboratory staff, participated in prolonged sessions of continuous reading aloud. The male subject read for 3.5 hours on one day, and the female subject read for four hours on one day and 3.5 hours the following day. For both subjects, recordings were made of the stimulus materials at the beginning and at the end of the 3.5 hour reading sessions. Between reading the stimulus materials, both subjects read aloud from a novel. To increase the likelihood of causing vocal fatigue, subjects tried to read at a level above 60 dB during the entire session. Vocal level was monitored by an experimenter.

For each subject, the target words from the start and end of the 3.5 hour reading session were analyzed and compared. The following acoustic features were identified and examined for each syllable of the target words: average amplitude of the vowel, amplitude variability within the vowel, average F0 of the vowel, F0 variability within the vowel, vowel duration, vowel/word duration ratio, vowel/sentence duration ratio, average F1 of the vowel, average F2 of the vowel, and average F3 of the vowel.

Of the parameters investigated, prolonged periods of talking had a significant effect only on measures of amplitude and F0. With prolonged speech, the average amplitude and the amplitude variability decreased, while the average F0 and F0 variability increased. No consistent pattern of changes was observed across conditions for either duration or formant frequencies of F1, F2, or F3. Additionally, there were no interactive effects between prolonged speech and the presence of lexical stress.

In addition to gross changes in F0, we also carried out four fine-grained analyses of the fundamental frequency. These included measures of pitch perturbation, amplitude perturbation (Davis, 1976), directional perturbation (Hecker and Kreul, 1971), and noise-to-harmonic ratio (Muta and Baer, 1988). The pitch and amplitude perturbation measures are algorithms that describe the magnitude of changes that occur in the fundamental frequency across conditions. The directional perturbation factor describes the increase or decrease of pitch period duration perturbations in a manner that is independent of the magnitude of the changes. The noise-to-harmonics ratio provides information related to the breathiness or hoarseness at the glottal source. The ratio is calculated by dividing the average energy of the lowest band between each harmonic of the fundamental by the average energy of the entire spectrum. A low noise-to-harmonics ratio is indicative of a hoarse voice.

The data showed large individual differences between our two subjects. Comparing across fatigue conditions, the pitch perturbation measure increased from normal to fatigued for our female talker, but decreased for our male talker. Both talkers showed a decrease in amplitude perturbation from the normal to fatigue conditions. DB, the female talker, showed a significant decrease for the vowel /ah/. KJ, the male talker, showed a significant decrease for the vowel /uh/. Confounding this result, however, is the fact that DB showed a significant increase in amplitude perturbation across conditions for the vowel /ae/. Neither subject showed a significant effect in

the noise-to-harmonics ratio.

Taken together, the results from this study indicate that it was difficult to induce vocal fatigue in our subjects. We believe that with a more controlled study, we might be better able to induce fatigue experimentally on demand. A more rigorous experiment would place high attentional demands on the subjects and would also require the subjects to be highly motivated throughout the entire testing session. Given the lack of any quantitative experimental research in the area of vocal fatigue, we believe that more research will be necessary to define what the salient acoustic-phonetic effects are and how they can be produced reliably in the laboratory.

Summary and General Conclusions

Our work under this contract has examined speech production and perception in severe environments. We have carried out three sets of experiments: the first set examined the Lombard effect, specifically the acoustic-phonetic properties of speech produced in noise, the second examined speech produced under cognitive load using an attention-demanding tracking task, and the third examined vocal fatigue.

In general, five factors emerged from these three projects. First, we observed fairly consistent changes in amplitude and spectral tilt across our three studies. Subjects under workload produced utterances that were louder and more monotone in pitch. Second, concurrent with changes in amplitude and spectral tilt were decreases in duration. Subjects tended to shorten their utterances when speaking under a physical or mental workload. A third factor that has become apparent to us is that two types of articulatory changes are involved in these effects. The first deals with acoustic changes that occur above the level of the larynx. For example, in the Lombard experiments, we observed fairly consistent increases in F1 while F2 remained constant. These changes in speech are indicative of supralaryngeal adjustments that occur under stress. The second type of change deals with modifications to the glottal source. These changes need to be addressed in both their absolute magnitude and in their relative changes on a pitch period to pitch period basis. Another factor that needs further study is the changes in the variability at the glottal source. Many of the changes we observed involved adjustments and modifications in the shape of the glottal waveform. The present findings demonstrate that not only do the absolute values of the acoustic measures change across conditions, but the variability of these measures changes also. The final factor deals with the wide variation we observed among individual subjects. Across all three sets of experiments, we have found that subjects often display different and conflicting changes in their speech.

These conclusions suggest two areas for future research. First, experimentation must be more rigorous in order to reduce the wide variations that we have observed among our subjects. In most of our studies we used only a small number of subjects. Second, individual differences must be taken into account when designing automated devices that would be used with a wide range of talkers. The research carried out under this contract demonstrates the importance of these changes in the acoustic-phonetic properties of speech in severe environments. These findings should be very useful in designing the

next generation of speech recognition systems that are expected to display extremely robust performance across a wide range of conditions.

The basic research carried out under this contract has attempted to fill a serious gap in our current knowledge base. Most of what we know about speech perception and production has been obtained in benign laboratory environments with extremely cooperative subjects who are not under any environmental or cognitive load. We are now beginning to realize that while this research literature is extremely useful for theoretical studies of speech, many of the findings do not generalize easily to more severe environments. The present findings point to the need for additional basic research on the acoustic-phonetic properties of speech across a wide range of environments and conditions. Understanding the nature of acoustic-phonetic variability in the speech signal is one of the most important problems that will need to be solved in order to develop robust speech recognition technologies for use in severe environments such as aircraft cockpits or air traffic control towers.

Staffing

Personnel 7/1/86 - 6/30/90

Name	Title	Months on Payroll	Percent Effort (Approximate)
David B. Pisoni	Principal Investigator	4	30 summer 10 acad. yr.
Jerry Forshee	Systems Analyst	37	< 10
David Link	Electronics Engineer	37	10
Robert Bernacki	Research Systems Engineer	18	20
Beth Greene	Assistant Scientist	20	< 10
Van Summers	Research Associate	19	60
Moshe Yuchtman	Research Associate	5	80
Luis Hernandez	Applications Programmer	18	25
Dennis Feaster	Applications Programmer	14	20
Michael Dedina	Applications Programmer	6	25
Michael Stokes	Graduate Assistant	30	25
Robert Pedlow	Graduate Assistant	18	25
Scott Lively	Graduate Assistant	12	25
Cheryl Blackerby	Secretary	18	15
Mary Stapleton	Secretary	9	20
Hourly	Clerical Assistants	1000 hours	

Contract Related Travel

1. The Principal Investigator travelled to Dayton, Ohio in October, 1986 for the meeting of the Human Factors Society.

2. The Principal Investigator attended the conference of the Association for Research in Otolaryngology in Clearwater, Florida during February of 1987.

3. Lab personnel went to Indianapolis, Indiana in May, 1987 for the Spring meeting of the Acoustical Society of America.

4. Lab personnel working on the Air Force project attended the Spoken Language Workshop in Buffalo, New York and then travelled to the meeting of the Acoustical Society of America in Syracuse, New York during May of 1989.

5. Lab personnel went of Dayton, Ohio in June, 1989 to visit Wright Patterson Air Force Base and to meet with the contract monitor, Dr. Thomas J. Moore.

Publications

Makhoul, J. (Chairman), Crystal, T., Green, D., Hogan, D., McAulay, D., Pisoni, D., Sorkin, R., Stockham, D. (1989). Removal of noise from noise-degraded speech signals. CHABA Report. Washington, DC: National Academy Press.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. Journal of the Acoustical Society of America, 84, 917-928.

Summers, W. V., Johnson, K. A., Pisoni, D. B., Bernacki, R. H. (1989). An addendum to "Effects of noise on speech production: Acoustic and perceptual analyses"[J.Acoust.Soc.Am. 84, 917-928 (1988)]. Journal of the America, 86 (5), 1717-1721.

Conference Presentations and Invited Talks

Pisoni, D. B. (1989). Development of new methods the assess spoken language comprehension. Paper presented at T&E Workshop, RADC/EEV, DDVPC, Bedford, MA.

Pisoni, D. B. (1990). Human factors in auditory perception. Paper presented at the Workshop on the Quality of Digitally Processed Acoustic Signals, Mystic, CT.

Summers, W. V., Pisoni, D. B., & Stokes, M. A. (1989). Effects of cognitive workload on speech production. Paper presented at the Spring meeting of the Acoustical Society of America, Syracuse, NY.

SRL Progress Report and Technical Note Series

Summers, W. V., Pisoni, D. B., & Bernacki, R. H. (1989). Effects of cognitive workload on speech production: Acoustic analyses. Research on Speech Perception in Progress: Progress Report No. 15. Indiana University. Pp. 485-502.

References

- Davis, S. B. (1976). Computer evaluation of laryngeal pathology based on inverse filtering of speech. SCRL Monograph, 13.
- Draegert, G. L. (1951). Relationships between voice variables and speech intelligibility in high level noise. Speech Monograph, 18, 272-278.
- Fitch, H. (1989). Comments on the effects of noise on speech production: Acoustic and perceptual analyses. Journal of the Acoustical Society of America, 86, 2017-2019.
- Hanley, T. D., & Steer, M. D. (1949). Effects of level of distracing noise upon speaking rate, duration and intensity. Journal of Speech and Hearing Disorders, 14, 363-368.
- Hansen, J. H. L. (1988). Analysis and compensation of stressed and noisy speech with applications to robust automatic recognition. Unpublished doctoral dissertation, Georgia Institute of Technology.
- Hecker, M. H. L., Stevens, K. M., von Bismarck, G., & Williams, C. E. (1968). Manifestations of task-induced stress in the acoustic signal. Journal of the Acoustical Society of America, 44, 993-1001.
- Hecker, M. H. L., & Kruei, E. J. (1971). Descriptions of the speech of patients with cancer of the vocal folds. Journal of the Acoustical Society of America, 52, 1238-1250.
- Jex, H. R., McDonnell, J. D., & Phatak, A. V. (1966). A 'critical' tracking task for manual control research. In N. Moray, (Ed.). Mental workload. NY: Plenum Press.
- Kuroda, I., Fujiwara, O., Okamura, N., & Utsuki, N. (1976). Methods for determining pilot stress through analysis of voice communication. Aviation, Space, and Environmental Medicine, 47, 528-533.
- Lane, H. L., Tranel, B., & Sison, C. (1970). Regulation of voice communication by sensory dynamics. Journal of the Acoustical Society of America, 47, 618-624.
- Lane, H. L., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. Journal of Speech and Hearing Research, 14, 677-709.
- Lombard, E. (1911). Le signe de l'elevation de la voix. Ann. Mal. Oreil. Larynx, 37, 101-119. (Cited in Lane & Tranel, 1971).
- Muta, H., & Baer, T. (1988). A pitch-synchronous analysis of hoarseness in running speech. Journal of the Acoustical Society of America, 52, 1238-1250.

- Pisoni, D. B., Bernacki, R. H., Nusbaum, H. C., & Yuchtman, M. (1985). Some acoustic-phonetic correlates of speech produced in noise. In Proceedings of the 1985 International Conference on Acoustics, Speech, and Signal Processing. Pp. 1581-1584.
- Sander, E. K., & Ripich, D. E. (1983). Vocal fatigue. Ann. Otol Rhinol Laryngol, 92, 141-145.
- Scherer, R. C., Titze, I. R., Raphael, B. N., Wood, R. P., Ramig, L. A., & Blager, R. F. (1987). Vocal fatigue in a trained and untrained voice user. In T. Baer and L. Sasaki, (Eds.). Laryngeal function in phonation and respiration. Boston: Little, Brown & Co.
- Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In A. Valdman, (Ed.). Emotions in personality and psychopathology (pp. 493-529). NY: Academic Press.
- Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. Journal of the Acoustical Society of America, 84, 917-928.
- Summers, W. V., Johnson, K. A., Pisoni, D. B., & Bernacki, R. H. (1989). An addendum to "Effects of noise on speech production: Acoustic and perceptual analyses" [J.Acoust.Soc.Am. 84, 917-928 (1988)]. Journal of Acoustical Society of America, 86 (5), 1717-1721.
- Tolkmitt, F. J., & Scherer, K. R. (1986). Effect of experimentally induced stress on vocal parameters. Journal of Experimental Psychology, 12, 302-312.
- Williams, C. E., & Stevens, K. N (1969) On determining the emotional state of pilots during flight: An exploratory study. Aerospace Medicine, 40, 1369-1372.
- Yanigahara, N. (1967). Significance of harmonic changes and noise components in hoarseness. Journal of Speech and Hearing Research, 10, 531-541.

Effects of noise on speech production: Acoustic and perceptual analyses

W. Van Summers, David B. Pisoni, Robert H. Bernacki, Robert I. Pedlow, and Michael A. Stokes,

Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405

(Received 2 March 1988; accepted for publication 9 May 1988)

Acoustical analyses were carried out on a set of utterances produced by two male speakers talking in quiet and in 80, 90, and 100 dB SPL of masking noise. In addition to replicating previous studies demonstrating increases in amplitude, duration, and vocal pitch while talking in noise, these analyses also found reliable differences in the formant frequencies and short-term spectra of vowels. Perceptual experiments were also conducted to assess the intelligibility of utterances produced in quiet and in noise when they were presented at equal S/N ratios for identification. In each experiment, utterances originally produced in noise were found to be more intelligible than utterances produced in the quiet. The results of the acoustic analyses showed clear and consistent differences in the acoustic-phonetic characteristics of speech produced in quiet versus noisy environments. Moreover, these acoustic differences produced reliable effects on intelligibility. The findings are discussed in terms of: (1) the nature of the acoustic changes that take place when speakers produce speech under adverse conditions such as noise, psychological stress, or high cognitive load; (2) the role of training and feedback in controlling and modifying a talker's speech to improve performance of current speech recognizers; and (3) the development of robust algorithms for recognition of speech in noise.

PACS numbers: 43.70.Fq, 43.70.Gr, 43.71.Bp

INTRODUCTION

It has been known for many years that a speaker will increase his/her vocal effort in the presence of a loud background noise. Informal observations confirm that people talk much louder in a noisy environment such as a subway, airplane, or cocktail party than in a quiet environment such as a library or doctor's office. This effect, known as the Lombard reflex, was first described by Etienne Lombard in 1911 and has attracted a moderate degree of attention by researchers over the years. The observation that speakers increase their vocal effort in the presence of noise in the environment suggests that speakers monitor their vocal output rather carefully when speaking. Apparently, speakers attempt to maintain a constant level of intelligibility in the face of degradation of the message by the environmental noise source and the corresponding decrease in auditory sidetone at their ears. Lane and his colleagues (Lane *et al.*, 1970; Lane and Tranel, 1971) have summarized much of the early literature on the Lombard effect and have tried to account for a wide range of findings reported in the literature. The interested reader is encouraged to read these reports for further background and interpretation.

Despite the extensive literature on the Lombard effect over the last 30 years, little, if any, data have been published reporting details of the acoustic-phonetic changes that take place when a speaker modifies his vocal output while speaking in the presence of noise. A number of studies have reported reliable changes in the prosodic characteristics of speech produced in noise. However, very few studies have examined changes in the spectral properties of speech produced in masking noise.

In an earlier study, Hanley and Steer (1949) found that in the presence of masking noise, speakers reduce their rate of speaking and increase the duration and intensity of their utterances. In another study, Draeger (1951) examined the relations between a large number of physical measures of voice quality and speech intelligibility in high levels of noise and found a similar pattern of results. His interest was focused primarily on factors that correlated with measures of speech intelligibility rather than on a description of the acoustic-phonetic changes that take place in the speaker's speech. In addition to changes in duration and intensity, Draeger reported increases in vocal pitch and changes in voice quality due to a shift in the harmonic structure. The change in harmonic structure was shown by a difference in intensity between the low- and high-frequency components. To obtain these measures, the speech was bandpass filtered to obtain estimates of the locations of the major concentrations of energy in the spectrum. Unfortunately, no measurements of the size of the effects were reported in this article.

In another study on the intelligibility of speech produced in noise, Dreher and O'Neill (1957) reported that, when presented at a constant speech-to-noise ratio, speech produced by a speaker with noise in his ears is more intelligible than speech produced in quiet. This result was observed for both isolated words and sentences. In each case, for the noise condition, a broadband random noise source was presented over the speaker's headphones during production.

Related findings have been reported by Ladefoged (1967, pp. 163-165) in an informal study designed to examine how eliminating auditory feedback affects a speaker's speech. Auditory feedback was eliminated by presenting a loud masking noise over headphones at an intensity level

that prevented the subject from hearing his/her voice even via bone conduction. Subjects read a prepared passage and also engaged in spontaneous conversation. According to Ladefoged, although subjects' speech remained intelligible, it became "very disorganized" by removal of auditory feedback through the presentation of masking noise. Of special interest to us was the observation by Ladefoged that the length and quality of many of the vowel sounds were affected quite considerably by the masking noise. Some sounds became more nasalized, others lost appropriate nasalization. Pitch increased and there appeared to be much less variability in the range of pitch. Ladefoged also noticed a striking alteration in voice quality brought about by the tightening of the muscles of the pharynx. These findings were summarized informally by Ladefoged in his book without reporting any quantitative data. To our knowledge, these results have never been published. Nonetheless, they are suggestive of a number of important changes that may take place when speakers are required to speak under conditions of high masking noise.

The Dreher and O'Neill (1957) results suggest that masking noise which does not eliminate auditory feedback to the subject may have a positive influence on speech intelligibility. On the other hand, the Ladefoged (1967) findings suggest that this may not be the case when environmental noise is so loud that all auditory feedback is eliminated.

The present investigation is concerned with the effects of masking noise on speech production. Our interest in this problem was stimulated, in part, by recent efforts of the Air Force to place speech recognition devices in noisy environments such as the cockpits of military aircraft. Although it is obvious that background noise poses a serious problem for the operation of any speech recognizer, the underlying reasons for this problem are not readily apparent at first glance. While extensive research efforts are currently being devoted to improving processing algorithms for speech recognition in noise, particularly algorithms for isolated speaker-dependent speech recognition, a great deal less interest has been devoted to examining the acoustic-phonetic changes that take place in the speech produced by talkers in high ambient noise environments. If it is the case, as suggested by the published literature, that speakers show reliable and systematic changes in their speech as the noise level at their ears increases, then it would be appropriate to examine these differences in some detail and to eventually incorporate an understanding of these factors into current and future algorithm development. Thus the problem of improving the performance of speech recognizers may not only be related to developing new methods of extracting the speech signal from the noise but may also require consideration of how speakers change their speech in noisy or adverse environments.

As noted earlier, a search through the literature on speech communication and acoustic-phonetics published over the last 40 years revealed a number of studies on the effects of noise on speech production and speech intelligibility. While changes in duration, intensity, and vocal pitch have been reported, and while changes in voice quality have been observed by a number of investigators, little is currently known about the changes that take place in the distribution

of spectral energy over time such as modifications in the patterns of vowel formant frequencies or in the short-term spectra of speech sounds produced in noise. The present investigation was aimed at specifying the gross acoustic-phonetic changes that take place when speech is produced under high levels of noise as might be encountered in an aircraft cockpit. We expected to find reliable changes in prosodic parameters such as amplitude, duration, and vocal pitch, which have previously been reported in the literature. We were also interested in various segmental measures related to changes in formant frequencies and in the distribution of spectral energy in the short-term spectra of various segments. These measures might reflect changes in the speaker's source function as well as the articulatory gestures used to implement various classes of speech sounds. In the present study, digital signal processing techniques were used to obtain quantitative measures of changes in the acoustic-phonetic characteristics of speech produced in quiet and in three ambient noise conditions. A second aspect of the study involved perceptual testing with these utterances to verify Dreher and O'Neill's earlier finding that speech produced in noise was more intelligible than speech produced in quiet when the two conditions were presented at equivalent S/N ratios (Dreher and O'Neill, 1957).

1. ACOUSTIC ANALYSES

A. Method

1. Subjects

Two male native English speakers (SC and MD) were recruited as subjects. SC was a graduate student in psychology and was paid \$5.00 for his participation. MD was a member of the laboratory staff and participated as part of his routine duties. Both speakers were naive to the purpose of the study and neither speaker reported a hearing or speech problem at the time of testing. Both speakers served for approximately 1 h.

2. Stimulus materials

Stimulus materials consisted of the 15 words in the Air Force speech recognition vocabulary: the digits "zero," "one," "two," "three," "four," "five," "six," "seven," "eight," "nine"; and the control words "enter," "frequency," "step," "threat," and "CCIP." These words were typed into computer files and different randomizations of the list of 15 words were printed out for the subjects to read during the course of the experiment.

3. Procedure

Subjects were run individually in a single-walled sound-attenuated booth (IAC model 401 A). The subject was seated comfortably in the booth and wore a pair of matched and calibrated TDH-39 headphones. An Electrovoice condenser microphone (model C090) was attached to the headset with an adjustable boom. Once adjusted, the microphone remained at a fixed distance of 4 in. from the subject's lips throughout the experiment.

The masking noise consisted of a broadband white noise source that was generated with a Grason-Stadler noise generator (model 1724). The noise was low-pass filtered at 3.5 kHz, using a set of Krohn-Hite filters (model 3202R) with a roll-off of 24 dB per octave, and passed through a set of adjustable attenuators. The masking noise was presented binaurally through the headphones. Subjects wore the headphones during the entire experiment.

Subjects read the words on the test lists under four conditions: quiet, 80, 90, or 100 dB of masking noise in their earphones. The quiet condition measured from 33- to 37-dB SPL background noise with the attenuators set to their maximum setting. Measurements of the noise were made with a B&K sound level meter and artificial ear connected to the earphones.

After the headset was adjusted and the subject became familiar with the environment, a sheet of written instructions was provided to explain the procedures that would be followed. Subjects were informed that they would be reading English words from a list and that they should say each word clearly with a pause of about 1–2 s between words. They were told that masking noise at various levels of intensity would be presented over their headphones during the course of the experiment and that their task was to read each word as clearly as possible into the microphone. They were also told that the experimenter would be listening to their speech outside the booth while the recording was being made. Before the actual recordings were made, both subjects were given about 15 min of practice reading lists of the vocabulary with no noise over the headphones. This was done to familiarize the subjects with the specific vocabulary and the general procedures to be used in making the audiotapes.

Data were collected from subjects reading the lists under all four noise conditions. The noise levels were randomized within each block of four lists with the restriction that over the entire experimental session, every noise level was followed by every other noise level except itself. Subjects took about 40 s to read each list. After each list was read, the masking noise was turned off for about 40 s during which the subjects sat in silence. Each list of 15 words was read in each of the four masking conditions five times, for a total of 300 responses from each subject. Recordings were made on an Ampex AG-500 tape recorder running at 7½ ips.

4. Speech signal processing

Productions of the digits "zero," "one," "two," "three," "four," "five," "six," "seven," "eight," and "nine" were analyzed using digital signal processing techniques. These 400 utterances (ten words \times five repetitions \times four noise levels \times two talkers) were digitized using a VAX 11/750 computer. The utterances were first low-pass filtered at 4.8 kHz and then sampled at a rate of 10 kHz using a 16-bit A/D converter (Digital Sound Corporation model 2000). Each utterance was then digitally edited using a cursor-controlled waveform editor and assigned a file name. These waveform files were then used as input to several digital signal processing analyses.

Linear predictive coding (LPC) analysis was per-

formed on each waveform file. LPC coefficients were calculated every 12.8 ms using the autocorrelation method with a 25.6-ms Hamming window. Fourteen linear prediction coefficients were used in the LPC analyses. The LPC coefficients were used to calculate the short-term spectrum and overall power level of each analysis frame (window). Formant frequencies, bandwidths, and amplitudes were also calculated for each frame from the LPC coefficients. In addition, a pitch extraction algorithm was employed to determine if a given frame was voiced or voiceless and, for voiced frames, to estimate the fundamental frequency (F_0).

Total duration for each utterance was determined by visual inspection and measurement from a CRT display that simultaneously presented the utterance waveform along with time-aligned, frame-by-frame plots of amplitude, F_0 (for voiced frames), and formant parameters. Cursor controls were used to locate the onset and offset of each utterance. Following identification of utterance boundaries, a program stored the total duration, mean F_0 , and mean rms energy for each utterance. The onset and offset of the initial vowel of each utterance were also identified and labeled. For each utterance, mean formant frequencies from this vowel segment were also stored. In the case of the word "zero," the initial vowel /i/ could not be reliably segmented apart from the following voiced segments; thus the entire /iro/ segment was used as the initial vowel for this utterance. Similarly, for the utterances "three" and "four," the semivowel /r/ was included as part of the vowel during segmentation.

Finally, the peak amplitude frame (25.6-ms window) from the stressed vowel of each utterance was identified and a regression line was fit to the spectrum of this analysis frame. The slope of this regression line was taken as a measure of "spectral tilt," to quantify the relative distribution of spectral energy at different frequencies.

B. Results and discussion

The influence of ambient noise on various acoustic characteristics of the test utterances is described below. In each case, an analysis of variance was used to determine whether noise level had a significant effect on a given acoustic measure. Separate analyses were carried out for the two talkers. The analyses used word ("zero" through "nine") and noise level as independent variables. The presentation of results will focus on the effect of noise on the various acoustic measures. The "word" variable will be discussed only in cases where a significant word \times noise interaction was observed.

1. Amplitude

Mean rms energy for utterances spoken at each noise level are shown for each talker in Fig. 1. The data are collapsed across utterances. For each talker, the measured amplitudes show a consistent increase with an increase in noise level at the talker's ears. The largest increase occurred between the quiet condition and the 80-dB noise condition. Analyses of variance revealed that, for each talker, noise level had a significant effect on amplitude [$F(3,160) = 190.41$, $p < 0.0001$ for talker MD, and $F(3,160) = 211.15$, $p < 0.0001$ for talker SC]. Newman-Keuls multiple range analyses revealed that, for each talker,

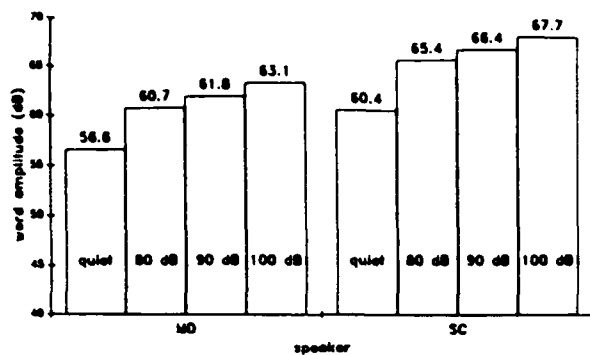


FIG. 1. Mean rms amplitudes for words produced in quiet, 80, 90, and 100 dB of masking noise. Values are collapsed across utterances and presented separately for each speaker.

each increase in noise led to a significant increase in amplitude (all p s < 0.01). For talker MD, there was also a significant word \times noise interaction [$F(27,160) = 1.72, p < 0.03$]. For both speakers, the pattern of increased masking noise producing an increase in amplitude was present for every word. The word \times noise interaction for speaker MD is due to variability across words in the amount of amplitude increase.

2. Duration

Mean word durations for utterances spoken at each noise level are shown for each speaker in Fig. 2. The data are again collapsed across utterances. The pattern is similar to that observed for amplitude: Word duration shows a consistent increase with each increase in noise at the speakers' ears. However, for speaker MD, the change in duration between the 80- and 90-dB conditions is very small (6 ms). For SC, there is only a slight (15-ms) change in duration across the 80-, 90-, and 100-dB noise conditions. Analyses of variance demonstrated that, for each speaker, noise had a significant effect on word duration [$F(3,160) = 23.08, p < 0.0001$ for speaker MD, and $F(3,160) = 25.31, p < 0.0001$ for speaker SC]. Newman-Keuls analyses revealed that, for speaker MD, word duration was significantly shorter in the quiet condition than in any of the other conditions (p s < 0.01),

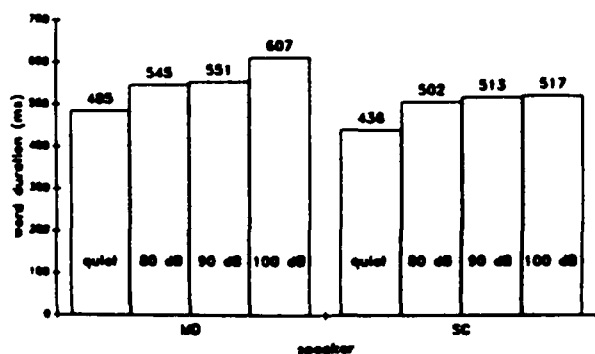


FIG. 2. Mean durations for words produced in quiet, 80, 90, and 100 dB of masking noise. Values are collapsed across utterances and presented separately for each speaker.

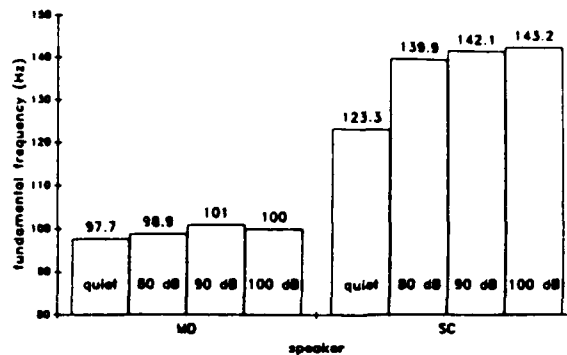


FIG. 3. Mean fundamental frequency values for words produced in quiet, 80, 90, and 100 dB of masking noise. Values are collapsed across utterances and presented separately for each speaker.

and significantly longer in the 100-dB condition than in the other conditions (p s < 0.01). Durations did not significantly differ in the 80- and 90-dB conditions for MD. For speaker SC, Newman-Keuls tests revealed that duration in the quiet condition was significantly shorter than in the other three conditions (p s < 0.01), but that duration did not significantly vary among the 80-, 90-, and 100-dB noise conditions.

3. Fundamental frequency

Mean fundamental frequencies for utterances spoken at each noise level are plotted separately for each speaker in Fig. 3. The data demonstrate a larger change in F_0 across noise conditions for speaker SC than for MD. For MD, F_0 showed a small increase as the noise increased from quiet to 80 dB to 90 dB, followed by a slight drop in F_0 between the 90- and 100-dB conditions. For SC, a large jump in F_0 occurred between the quiet and 80-dB noise conditions, followed by small additional increases in F_0 in the 90- and 100-dB conditions. Analyses of variance showed a significant effect of noise on F_0 for each speaker [$F(3,160) = 3.53, p < 0.02$ for speaker MD, and $F(3,160) = 42.07, p < 0.0001$ for speaker SC]. Newman-Keuls analyses revealed a significant change in F_0 between the quiet and 90-dB condition for speaker MD ($p < 0.05$). For speaker SC, the Newman-Keuls tests showed that F_0 in the quiet condition was significantly lower than in any of the other noise conditions (p s < 0.01).

4. Spectral tilt

As mentioned earlier, a regression line was fit to the spectrum of a representative frame from each token. The peak amplitude frame from the initial vowel was identified and used for these measurements. The slope of the regression line was taken as a measure of "spectral tilt" to index the relative energy at high versus low frequencies. Mean spectral tilt values for utterances spoken at each noise level are plotted for each speaker in Fig. 4. For each speaker, there was a decrease in spectral tilt accompanying each increase in noise. This decrease in tilt reflects a change in the relative distribution of spectral energy so that a greater proportion of energy is located in the high-frequency end of the spectrum when utterances are produced in noise. Analyses of variance demonstrated a significant change in spectral tilt across noise

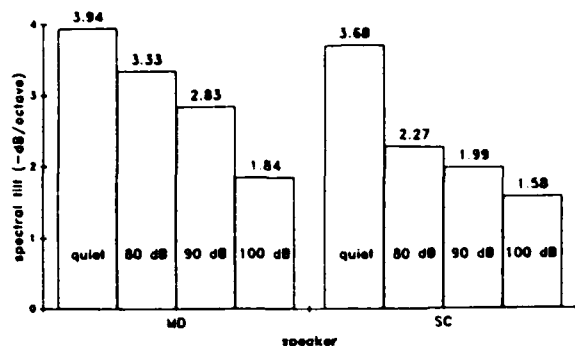


FIG. 4. Mean spectral tilt values for words produced in quiet, 80, 90, and 100 dB of masking noise. Values are collapsed across utterances and presented separately for each speaker.

conditions for each speaker [$F(3,160) = 56.82$, $p < 0.0001$ for speaker MD, and $F(3,160) = 23.85$, $p < 0.0001$ for speaker SC]. Newman-Keuls analyses revealed a significant decrease in spectral tilt with each increase in noise for speaker MD ($ps < 0.01$). For SC, spectral tilt was significantly greater in the quiet condition than in any of the other noise conditions ($ps < 0.01$). In addition, tilt was significantly greater in the 80-dB noise condition than in the 100-dB condition ($p < 0.05$).

On first examination, it appears that the decrease in spectral tilt observed in the high-noise conditions may be due to the increases in F_0 also observed in these conditions. However, a close examination of these two sets of results suggests that the relative increase in spectral energy at high frequencies in the high-noise conditions is not entirely due to increases in F_0 . For speaker MD, F_0 did not change a great deal across noise conditions (see Fig. 3); the change in F_0 was significant only in the quiet versus 90-dB comparison. Yet each increase in noise led to a significant decrease in spectral tilt for speaker MD. For speaker SC, the 80- and 100-dB noise conditions did not differ in the analysis of F_0 , yet a significant decrease in spectral tilt was obtained between these two conditions.

5. Formant frequencies

The influence of masking noise level on vowel formant frequencies was analyzed next. Mean F_1 and F_2 frequencies from the initial vowel of each utterance were examined. Noise had a consistent effect on the formant data for speaker SC and a less consistent effect for speaker MD. Mean F_1 frequencies for utterances produced in each noise condition are shown in Fig. 5. The data for speaker MD appear in the left-hand portion of the figure and the data for speaker SC appear in the middle of the figure. A significant main effect of noise on F_1 frequency was observed for speaker SC [$F(3,160) = 14.91$, $p < 0.0001$], along with a marginally significant noise \times word interaction [$F(27,160) = 1.5$, $p < 0.07$]. For this speaker, F_1 frequency tended to increase as the noise level increased. Newman-Keuls tests revealed that, for this speaker, F_1 was significantly lower in the quiet condition than in any of the other noise conditions. The marginally significant noise \times word interaction for SC suggests that the pattern of an increase in F_1 accompanying an increase in noise may not hold for all ten utterances. The con-

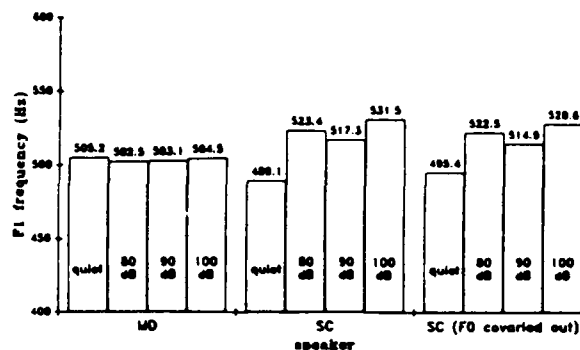


FIG. 5. Mean first formant frequency values for words produced in quiet, 80, 90, and 100 dB of masking noise. Values are collapsed across utterances and presented separately for speaker MD, speaker SC, and speaker SC with F_0 covaried out.

sistency of this pattern can be seen by examining Fig. 6. This figure displays F_1 and F_2 frequency data for the quiet and 100-dB noise conditions for each of the ten utterances produced by SC. With the exception of the utterance "one," F_1 was greater in the 100-dB condition than in the quiet condition for all utterances.

The main effect of noise and the noise \times word interaction did not reach significance in the analysis of F_1 frequency for speaker MD. As Fig. 5 shows, the change in mean F_1 frequency across noise conditions was less than 3 Hz for this speaker. The F_1 and F_2 data for speaker MD are broken down by utterance in Fig. 7. Although the noise \times word interaction was not significant for MD, the pattern of results shown in this figure suggests that the presence of masking noise may have produced a compacting, or reduction, in the range of F_1 for this speaker. In the majority of cases, utterances with low F_1 frequencies showed an increase in F_1 in noise, while utterances with high F_1 frequencies showed a decrease in F_1 .

The mean values shown in Figs. 3 and 5 demonstrate a striking similarity between the F_0 data and the F_1 data for each speaker. For speaker MD, there was little change in F_0 across noise conditions and no significant influence of noise on F_1 frequency. For speaker SC, both F_0 and F_1 were significantly higher in the 80-, 90-, and 100-dB noise conditions than in the quiet condition. These data suggest a close relationship between F_0 and F_1 ; apparently, an increase in fundamental frequency leads to an increase in F_1 . We carried out one additional analysis to further test this conclusion.

In order to determine whether F_0 and F_1 were, in fact, directly related, a second analysis was run on speaker SC's data. In this analysis, the effects of word and noise level on initial-vowel F_1 frequency were again tested but with initial-vowel F_0 entered as a covariate in the analysis. Mean F_1 frequencies at each noise level based on the adjusted cell means from this analysis (in which F_0 is covaried out) appear in the right-hand portion of Fig. 5. The results of this analysis were nearly identical to those observed in the original analysis of F_1 frequency for SC. The main effect of noise on F_1 frequency remained significant [$F(3,159) = 5.32$, $p < 0.0017$]. Also, as in the original analysis of F_1 for speaker SC, the noise \times word interaction fell short of significance [$F(27,159) = 1.44$, $p < 0.09$]. Finally, Newman-Keuls tests comparing F_1 frequencies in the various noise condi-

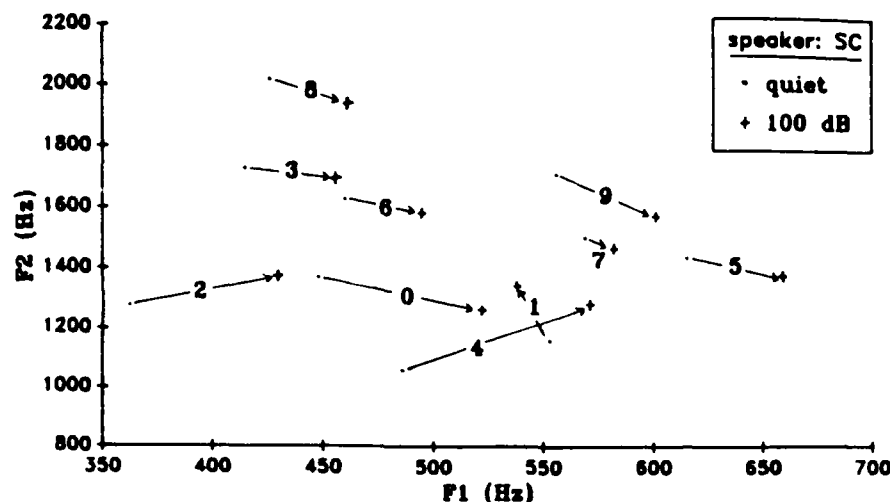


FIG. 6. Mean first and second formant frequencies for words produced in quiet and 100 dB of masking noise by speaker SC. Values are presented separately for each utterance.

tions revealed the identical pattern observed in the original analysis: $F1$ frequency was significantly lower in the quiet condition than in any of the other noise conditions. Thus, for speaker SC, it appears that noise had an influence on $F1$ frequency independent of its influence on $F0$.

Turning to the $F2$ data, masking noise did not produce a significant main effect on $F2$ frequency for speaker SC. However, a significant noise \times word interaction was present [$F(27,160) = 1.92, p < 0.008$]. An examination of Fig. 6 suggests that the range of $F2$ frequencies was reduced in the presence of noise for speaker SC. Utterances containing high $F2$ frequencies showed a decrease in $F2$ in the 100-dB condition, while utterances with low $F2$ frequencies showed increases in $F2$ when noise was increased.

The main effect of noise and the noise \times word interaction did not approach significance in the analysis of $F2$ frequency for speaker MD. An examination of Fig. 7 shows that, for most utterances, $F2$ showed little change between the quiet and 100-dB noise condition for this speaker.

Fundamental frequency, amplitude, and duration all tended to increase in the presence of noise. In addition, the results demonstrated consistent differences in the spectral characteristics of vowels produced in noise versus quiet. Vowels from utterances produced in noise had relatively flat

spectra, with a relatively large proportion of their total energy occurring in higher frequency regions. Vowels from utterances produced in quiet had steeper spectra with relatively little energy present in high-frequency regions. First formant frequencies also appeared to be influenced by the presence of noise for at least one speaker. For SC, $F1$ frequencies were higher for vowels from utterances produced in the three noise conditions than for vowels produced in the quiet. There was little change in $F2$ frequencies across noise conditions for either speaker.

The present results demonstrated several clear differences in the acoustic characteristics of speech produced in quiet compared to speech produced in noise. Previous research by Dreher and O'Neill (1957) suggests that the changes in the spectral and temporal properties of speech which accompany the Lombard effect improve speech intelligibility. We carried out two separate perceptual experiments to verify their earlier conclusions.

II. PERCEPTUAL ANALYSES—EXPERIMENT I

In experiment I, subjects identified utterances from the quiet condition and the 90-dB masking noise condition in a forced-choice identification task. Utterances from the quiet and 90-dB noise condition were mixed with broadband noise

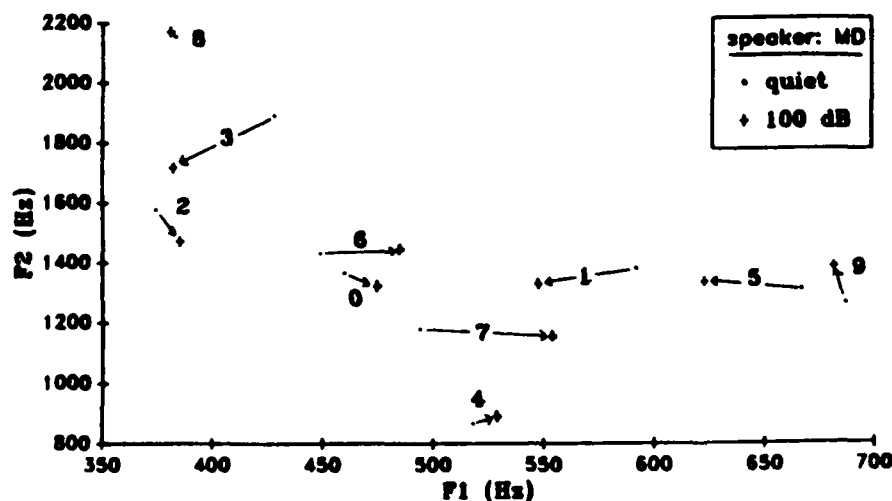


FIG. 7. Mean first and second formant frequencies for words produced in quiet and 100 dB of masking noise by speaker MD. Values are presented separately for each utterance.

at equivalent S/N ratios and presented to listeners for identification. If Dreher and O'Neill's conclusion concerning the intelligibility of speech produced in noise versus quiet is correct, subjects should identify utterances produced in the 90-dB noise condition more accurately than utterances produced in the quiet condition.

A. Method

1. Subjects

Subjects were 41 undergraduate students who participated to fulfill a requirement for an introductory Psychology course. All subjects were native English speakers and reported no previous history of a speech or hearing disorder at the time of testing.

2. Stimuli

Stimulus materials were the tokens of the digits zero through nine, produced in quiet and 90 dB of masking noise by both talkers. For each talker, the five tokens of each word produced in each masking condition were used for a total of 100 utterances per masking condition (five tokens \times ten digits \times two talkers). All stimuli were equated in terms of overall rms amplitude using a program that permits the user to manipulate signal amplitudes digitally (Bernacki, 1981).

3. Procedure

Stimulus presentation and data collection were controlled by a PDP 11/34 computer. Stimuli were presented via a 12-bit digital-to-analog converter over matched and calibrated TDH-39 headphones. Wideband noise, filtered at 4.8 kHz, was mixed with the signal during stimulus presentation.

The 200 utterances from the quiet and 90-dB masking conditions were randomized and presented to subjects in one of three S/N conditions: -5 -, -10 -, and -15 -dB S/N ratio. The S/N ratio was manipulated by varying signal amplitude while holding masking noise constant at 85 dB SPL. Stimuli were presented at 70 dB SPL in the -15 -dB S/N condition, 75 dB SPL in the -10 -dB S/N condition, and 80 dB SPL in the -5 -dB S/N condition.

Subjects were tested in small groups in a sound-treated room and were seated at individual testing booths equipped with terminals interfaced to the PDP 11/34 computer. At the beginning of an experimental trial, the message "READY FOR NEXT WORD" appeared on each subject's terminal screen. The 85-dB SPL masking noise was presented over the headphones 1 s later. A randomly selected stimulus was presented for identification 100 ms following the onset of masking noise. Masking noise was terminated 100 ms following stimulus offset. A message was then displayed on each subject's screen instructing the subject to identify the stimulus. Subjects responded by depressing one of the ten digit keys on the terminal keyboard. Subjects were presented with two blocks of 200 experimental trials. Within each block, each of the 200 utterances was presented once.

4. Design

All 200 test utterances were presented to each subject. Thus talker (MD or SC) and masking noise condition (quiet or 90 dB) were manipulated as within-subjects factors. The S/N ratio in the listening conditions was manipulated as a between-subjects factor. Subjects were randomly assigned to one of the three S/N conditions. Thirteen subjects participated in the -15 -dB S/N condition, 13 participated in the -10 -dB condition, and 15 participated in the -5 -dB condition.

B. Results

The percentage of correct digit responses is displayed separately by speaker (MD or SC), masking noise condition (quiet or 90 dB SPL), and S/N ratio (-5 -, -10 -, or -15 dB) in Fig. 8. A three-way ANOVA was carried out on these data using speaker, masking noise, and S/N ratio as independent variables.

As expected, S/N ratio had a significant main effect on identification [$F(2,38) = 202.91, p < 0.0001$]. As shown in Fig. 8, performance was highest in the -5 -dB S/N condition, somewhat lower in the -10 -dB condition, and lowest in the -15 -dB condition. This pattern was observed for both talkers and for both the quiet and 90-dB noise conditions.

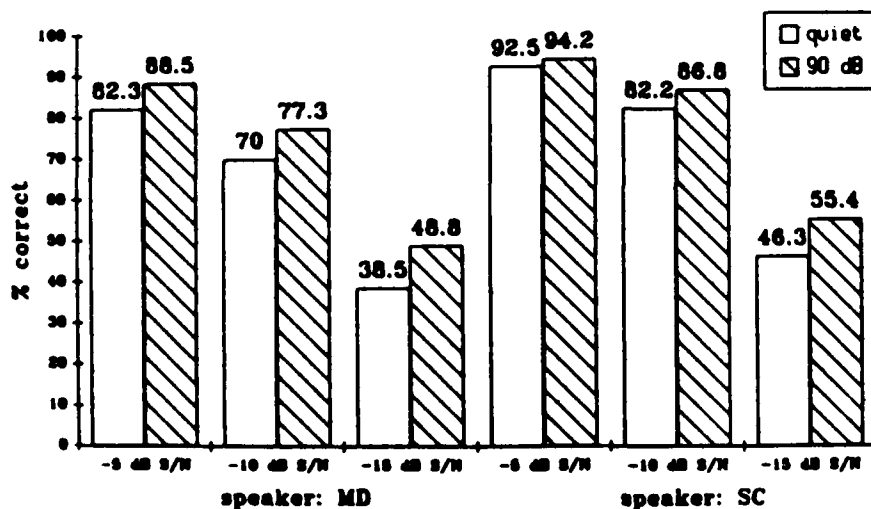


FIG. 8. Intelligibility of words produced in quiet and 90 dB of masking noise (experiment 1). Performance is broken down by S/N ratio and speaker.

Turning to the main focus of the experiment, masking noise produced a significant main effect on identification [$F(1,38) = 162.75, p < 0.0001$]. Digits produced in 90 dB of masking noise were consistently identified more accurately than digits produced in the quiet regardless of talker or S/N ratio (see Fig. 8).

A significant interaction was observed between masking noise and S/N ratio [$F(2,38) = 11.04, p < 0.0003$]. For each speaker, as S/N ratio decreased, the effect of masking noise on identification accuracy increased. Thus the difference in performance for digits produced in quiet versus 90-dB masking noise was smallest in the -5-dB S/N condition, greater in the -10-dB condition, and greatest in the -15-dB condition. Apparently, the acoustic-phonetic differences between the utterances produced in quiet versus 90 dB of noise had a greater influence on intelligibility as the S/N ratio decreased.

A significant noise \times talker interaction was also obtained [$F(1,38) = 5.68, p < 0.03$]. At each S/N ratio, the influence of masking noise on identification accuracy was greater for speaker MD than for speaker SC.

III. PERCEPTUAL ANALYSES—EXPERIMENT II

In experiment I, digits produced in noise were recognized more accurately than digits produced in quiet. The consistency of this effect in experiment I is quite remarkable given that the stimuli were drawn from a very small, closed set of highly familiar test items. To verify that the results of experiment I were reliable and could be generalized, we replicated the experiment with a different set of stimuli drawn from the original test utterances.

A. Method

Experiment II was carried out with stimuli taken from the 100-dB masking condition. That is, the replication used the 200 stimuli from the quiet and 100-dB masking conditions. In this experiment, ten subjects participated in the -15-dB S/N condition, nine subjects participated in the -10-dB condition, and ten subjects participated in the -5-dB condition. All subjects were native speakers of American English and met the same requirements as those

used in the previous experiment. All other aspects of the experimental procedure were identical to those of experiment I.

B. Results and discussion

The results of experiment II are shown in Fig. 9. Percent correct identification is broken down by speaker (MD or SC), masking noise (quiet or 100 dB SPL), and S/N ratio (-5, -10, or -15 dB). As in the previous experiment, a three-way ANOVA was carried out on these data using talker, masking noise, and S/N ratio as independent variables.

Comparing the data shown in Figs. 8 and 9, it can be seen that the pattern of means obtained in the two experiments is nearly identical. The results of the ANOVA performed on the data from this experiment also replicate the results of the previous experiment. A significant main effect of S/N ratio was obtained. Identification accuracy decreased as S/N ratio decreased [$F(2,26) = 117.33, p < 0.0001$]. There was also a significant main effect of talker [$F(1,26) = 39.79, p < 0.0001$]. As in the first experiment, SC's tokens were identified more accurately than MD's tokens.

Each of the significant effects involving the masking noise variable reported in experiment I was also replicated. There was a significant main effect of masking noise [$F(1,26) = 249.84, p < 0.0001$]. Utterances produced in 100 dB of noise were more accurately identified than utterances produced in the quiet. Significant interactions were observed between masking noise and S/N ratio [$F(2,26) = 9.46, p = 0.0009$] and between masking noise and talker [$F(1,26) = 41.16, p < 0.0001$]. As in experiment I, the effect of masking noise on identification accuracy increased as S/N ratio decreased. Also replicating the results of experiment I, the effect of masking noise on performance was greater for talker MD than for talker SC.

The results of these perceptual experiments replicate the findings of Dreher and O'Neill (1957). In their earlier research, as in each of the perceptual experiments reported here, subjects were more accurate at identifying utterances originally produced in noise than utterances produced in quiet. This pattern was found for each talker's utterances and at each S/N ratio in the present experiments. Further-

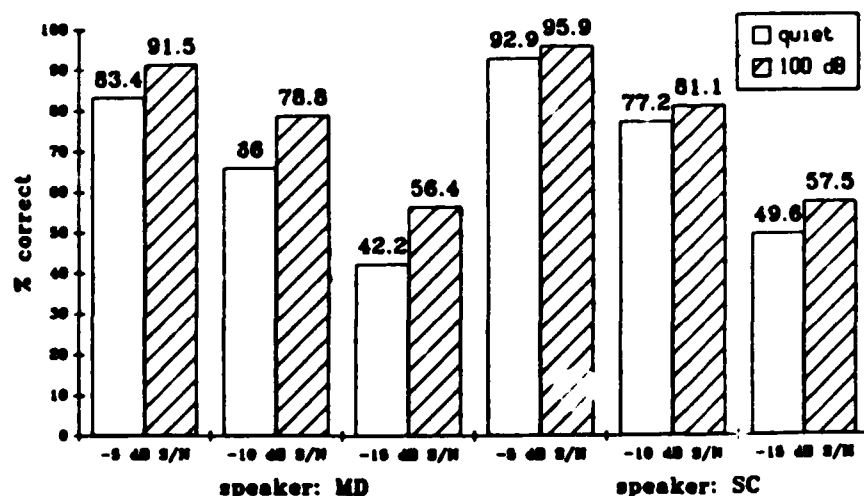


FIG. 9. Intelligibility of words produced in quiet and 100 dB of masking noise (experiment II). Performance is broken down by S/N ratio and speaker.

more, in each experiment, the effect of masking noise on intelligibility increased as S/N ratio decreased. Thus differences in the acoustic-phonetic structure of utterances produced in quiet and utterances produced in noise had reliable effects on intelligibility. The magnitude of these effects increased as the environment became more severe (as S/N ratio decreased).

IV. GENERAL DISCUSSION

The results of the present acoustic analyses demonstrate reliable and consistent differences in the acoustic properties of speech produced in quiet environments and environments containing high levels of masking noise. The differences we observed in our analyses were not restricted only to the prosodic properties of speech such as amplitude, duration, and pitch, but were also present in measurements of vowel formant frequencies. Moreover, for both talkers, we observed substantial changes in the slopes of the short-term power spectra of vowels in these utterances, which were shifted upward to emphasize higher frequency components.

The changes in amplitude, fundamental frequency, and duration reported here were often fairly small across the different noise levels. In particular, in comparing the 80-dB and 100-dB conditions, the change in amplitude was about 2 dB for each speaker. This 2-dB increase in the face of a 20-dB increase in masking noise is much smaller than would be predicted from previous research. Research using communication tasks involving talker-listener pairs have generally reported a 5-dB increase in signal amplitude for each 10-dB increase in noise (Lane *et al.*, 1970; Webster and Klump, 1962). The smaller differences observed in the present study suggest that masking noise may have a greater influence on speech in interactive communication tasks involving talker-listener pairs than in noninteractive tasks, such as the one used here, where no external feedback is available. Despite the magnitude of the observed differences, the findings are reliable and demonstrate important changes in speech produced in various noise conditions.

The results from the two perceptual experiments demonstrated that speech produced in noise was more intelligible than speech produced in quiet when presented at equal S/N ratios. Apparently, several acoustic characteristics of speech produced in noise, above and beyond changes in rms amplitude, make it more intelligible in a noisy environment than speech produced in the quiet. The present results also show that these acoustic differences play a greater and greater role as the S/N ratio decreases in the listener's environment.

The present findings replicate several gross changes in the prosodic properties of speech which have been previously reported in the literature (Hanley and Steer, 1949; Draeger, 1951). For one of our two speakers, the results also demonstrate a clear influence of masking noise on the formant structure of vowels. We believe that the present results have a number of important implications for the use of speech recognition devices in noisy environments and for the development of speech recognition algorithms, especially algorithms designed to operate in noise or severe environments.

In the recent past, a major goal of many speech scientists and engineers working on algorithm development has been to improve recognition of speech in noise (Rollins and Wiesen, 1983). Most efforts along these lines have involved the development of efficient techniques to extract speech signals from background noise (Neben *et al.*, 1983). Once the speech signal was extracted and the noise "stripped off" or attenuated, recognition could proceed via traditional pattern recognition techniques using template matching. Other efforts have attempted to solve the speech-in-noise problem by developing procedures that incorporate noise into the templates that is similar to the noise in the testing environment (Kersteen, 1982). By this technique, the signal does not have to be extracted from the noise; rather the entire pattern containing signal and noise is matched against the stored template.

This second technique, of incorporating noise into the templates, is accomplished by training the speech recognizer in a noisy environment so that noise along with speech is sampled on each trial. Kersteen (1982) reported success with this method of training; the highest recognition performance was produced when training and testing occurred in the same noise environment. Kersteen (1982) interpreted these results as demonstrating the importance of incorporating noise into the templates when noise is also present at testing.

An alternative explanation for the success of this training method is that the templates produced in noise capture acoustic characteristics of speech produced in noise that differ from those of speech produced in quiet. Unfortunately, little, if any, attention has been devoted to examining the changes in the speech signal that occur when a talker speaks in the presence of masking noise. The present findings demonstrate reliable differences in the acoustic-phonetic structure of speech produced in quiet versus noisy environments. Because of these differences, the problem of speech recognition in noise is more complicated than it might seem at first glance. The problem involves not only the task of identifying what portion of the signal is speech and what portion is noise but it also involves dealing with the changes and modifications that take place in the speech signal itself when the talker produces speech in noise.

Any speech recognition algorithm that treats speech as an arbitrary signal and fails to consider the internal acoustic-phonetic specification of words will have difficulty in recognizing speech produced in noise. This difficulty should be particularly noticeable with the class of current algorithms that is designed around standard template matching techniques. These algorithms are, in principle, incapable of recovering or operating on the internal acoustic-phonetic segmental structure of words and the underlying fine-grained spectral changes that specify the phonetic feature composition of the segments of the utterance. Even if dynamic programming algorithms are used to perform time warping before pattern matching takes place, the problems we are referring to here still remain. Factors such as changes in speaking rate, masking noise, or increases in cognitive load may affect not only the fairly gross attributes of the speech signal but also the fine-grained segmental structure

as well. Moreover, as far as we can tell, changes in speaking rate, effects of noise, and differences in cognitive load, to name just a few factors, appear to introduce nonlinear changes in the acoustic-phonetic realization of the speech signal. To take one example, it is a well-known finding in the acoustic-phonetic literature that consonant and vowel durations in an utterance are not increased or decreased uniformly when a talker's speaking rate is changed (see Miller, 1981, for a review). Thus simple linear scaling of the speech will not be sufficient to capture rate-related changes in the acoustic-phonetic structure.

The present findings are also relevant to a number of human factors problems in speech recognition. Both of the speakers examined in this study adjusted their speech productions in response to increased masking noise in their environment. These adjustments made the speech produced in noise more intelligible than speech produced in quiet when both were presented at equal amplitudes in a noisy environment. The speakers appeared to automatically adjust the characteristics of their speech to maintain intelligibility without having been explicitly instructed to do so. Presumably, the increase in intelligibility would have been at least as great if the speakers had been given such instructions. Given the currently available recognition technology, it should be possible to train human talkers to improve their performance with speech recognizers by appropriate feedback and explicit instructions.

In this regard, the present findings are related to several recent investigations in which subjects received explicit instructions to speak clearly (Chen, 1980; Picheny *et al.*, 1985; Picheny *et al.*, 1986). These studies on "clear speech" suggest that subjects can readily adjust and modify the acoustic-phonetic characteristics of their speech in order to increase intelligibility. Picheny *et al.* (1985) collected nonsense utterances spoken in "conversational" or "clear speech" mode. In conversational mode, each talker was instructed to produce the materials "in the manner in which he spoke in ordinary conversation." In clear speech mode, talkers were instructed to speak "as clearly as possible." Utterances produced in clear speech mode were significantly more intelligible than utterances produced in conversational mode when presented to listeners with sensorineural hearing losses. Chen (1980) reported the same pattern of results when "clear" and "conversational" speech was presented to normal-hearing subjects in masking noise.

Picheny *et al.* (1986) and Chen (1980) also carried out acoustic analyses to identify differences in the acoustic characteristics of clear and conversational speech. Many of the differences they identified are similar to those reported here. Specifically, longer segment durations, higher rms amplitudes, and higher F_0 values were reported for clear speech versus conversational speech. These changes in amplitude, duration, and pitch are also characteristic of speech that is deliberately emphasized or stressed by the talker (Lieberman, 1960; Klatt, 1975; Cooper *et al.*, 1985). Thus clear speech, emphasized or stressed speech, and speech produced in noise all tend to show increases in these three prosodic characteristics.

The pattern of formant data shows less similarity be-

tween speech produced in noise and clear speech or emphasized (stressed) speech. Chen (1980) reported that in clear speech F_1 and F_2 moved closer to target values. This movement enlarges the vowel space and makes formant values for different vowels more distinct, a pattern that is also characteristic of stressed vowels (Delattre, 1969). Our vowel formant data do not display this pattern of change. In the present study, masking noise produced increases in F_1 frequency for speaker SC but had little effect on formant frequencies for MD. Thus it appears that the presence of masking noise did not produce the same qualitative changes in production as instructions to speak clearly or to stress certain utterances. While several parallel changes occur in each case, a number of differences are also present in the data.

The literature on "shouted" speech also provides an interesting parallel to the present findings. Increases in fundamental frequency, vowel duration, and F_1 frequency have all been reported for shouted speech (Rostolland and Parant, 1974; Rostolland, 1982a,b). In addition, spectral tilt is reduced in shouted speech (Rostolland, 1982a). Each of these findings is in agreement with the present data for speech produced in noise. Thus it appears that the differences between speech produced in quiet and speech produced in noise are similar in kind to the differences between spoken and shouted speech. However, for each of the variables mentioned above, the differences between shouted and spoken speech are greater in magnitude than the present differences between speech produced in quiet and speech produced in noise.

In the present investigation, we found that speech produced in noise was more intelligible than speech produced in quiet when presented at equal S/N ratios. It would, therefore, be reasonable to expect that shouted speech should also be more intelligible than conversational speech in similar circumstances. However, the literature reports exactly the opposite result: When presented at equal S/N ratios, shouted speech is less intelligible than conversational speech (Pickett, 1956; Pollack and Pickett, 1958; Rostolland, 1985). While our talkers were able to increase the intelligibility of their speech by making changes in speech production that appear similar in kind to those reported for shouted speech, the magnitude of these changes is much greater in shouted speech. The extreme articulations that occur in shouted speech apparently affect intelligibility adversely, perhaps introducing distortions or perturbations in the acoustic realizations of utterances (Rostolland, 1982a,b).

In addition to the recent work of Picheny *et al.* (1985, 1986) and Chen (1980) on clear speech, there is an extensive literature in the field of speech communication from the 1940s and 1950s that was designed to improve the intelligibility of speech transmitted over noisy communication channels. Instructions to talk loudly, articulate more precisely, and talk more slowly have been shown to produce reliable gains in speech intelligibility scores when speech produced under adverse or noisy conditions is presented to human listeners for perceptual testing (see, for example, Tolhurst, 1954, 1955). Unfortunately, at the present time, we simply do not know whether these same training and feedback techniques will produce comparable improvements in perfor-

mance with speech recognition systems. It is clearly of some interest and potential importance to examine these factors under laboratory conditions using both speech recognizers and human observers. This line of research may yield important new information about the variability of different talkers and the "goat" and "sheep" problem discussed by Doddington and Schalk (1981). If we knew more precisely which acoustic-phonetic characteristics of speech spectra separate goats from sheep, we would be in a better position to suggest methods to selectively modify the way talkers speak to speech recognizers through training and directed feedback (see Nusbaum and Pisoni, 1987). We consider this to be an important research problem that has been seriously neglected by engineers and speech scientists working on the development of new algorithms for speech recognition. The human talker is probably the most easily modified component of a speech recognition system. In addition to being the least expensive component to change or modify, it is also the most accessible part of the system. Thus substantial gains in performance in restricted task environments should be observed simply by giving talkers directed feedback about precisely how they should modify the way they talk to the system. To this end, during training, the recognition system could provide the talker with much more information than a simple yes/no decision about the acceptance or rejection of an utterance. There is every reason to believe that a talker's speech can be modified and controlled in ways that will improve the performance of speech recognizers, even poorly designed recognizers that use fairly crude template-matching techniques.

We should qualify these remarks by also noting that these expected gains in performance can only be realized by additional basic research on how humans talk to speech recognizers under a wider variety of conditions. The results reported in the present article demonstrate that talking in the presence of masking noise not only affects the prosodic aspects of speech but also the relative distribution of energy across the frequency spectrum and the fine-grained acoustic-phonetic structure of speech as revealed in the formant frequency data. If we knew more about the role of feedback in speech production, and if we had more knowledge about the basic perceptual mechanisms used in speech perception, we would obviously have a much stronger and more principled theoretical basis for developing improved speech recognition algorithms specifically designed around general principles known to affect the way humans speak and listen.

The present investigation has a number of limitations that are worth discussing in terms of generalizing the findings beyond the present experimental context. First, we used isolated words spoken in citation form. It is very likely that a number of additional and quite different problems would be encountered if connected or continuous speech were used for these tests. The effects we observed with isolated words may be even more pronounced if the test words are put into context or concatenated together into short phrases or sentences.

Second, the subjects in this experiment did not receive any feedback about the success or failure of their communication. They were simply told that the experimenter was

listening and recording their utterances. Clearly, there was little incentive for the speaker to consciously change his speech even with masking noise present in the headphones. It seems reasonable to suppose that much larger changes might have been observed in the acoustic-phonetic properties of the utterances produced under masking noise if some form of feedback were provided to the talker to induce him to modify his articulations to improve intelligibility.

Finally, in these tests, no sidetone was provided to the talker through his headphones. In standard military communication tasks, sidetone is typically provided through the headphones and often serves as an important source of feedback that can modify the talker's speech output. As in the case of masking noise, it is not clear how auditory sidetone affects the acoustic-phonetic properties of talker's speech other than increasing or decreasing amplitude (see Lane and Tranel, 1971). Obviously, this is an area worthy of future research as it may be directly relevant to problems encountered in attempting to modify the way a talker speaks to a speech recognizer under adverse conditions. Thus automatic changes in the level of the sidetone may not only cause the talker to speak more loudly into the recognizer, but may also help him to articulate his speech more precisely and, therefore, improve performance with little additional cost to the system.

V. CONCLUSIONS

The problem of recognizing speech produced in noise is not just a simple problem of detection and recognition of signals mixed in noise. Speakers modify both the prosodic and segmental acoustic-phonetic properties of their speech when they talk in noise. Consequently, important changes in the physical properties of the speech signal must be considered along with the simple addition of noise to the signal in solving the recognition problem.

The presence of masking noise in a talker's ears not only affects the prosodic attributes of speech signals but affects the segmental properties as well. Talkers not only speak louder and slower in noise, but they also raise their vocal pitch and introduce changes in the short-term power spectrum of voiced segments. Talkers also introduce changes in the pattern of vowel formant frequencies.

In trying to articulate speech more precisely under these adverse conditions, the talker introduces certain changes in the acoustic-phonetic correlates of speech that are similar to those distinguishing stressed utterances from unstressed utterances. The changes in the prosodic properties of speech which occur in noise are also similar to changes that occur when subjects are explicitly instructed to "speak clearly." However, the F_1 and F_2 data suggest that the changes in productions that subjects automatically make when speaking in noise are not identical to the changes that occur when subjects are given clear speech instructions or when subjects put stress or emphasis on particular utterances.

The results of this study, taken together with the earlier findings reported in the literature on improving the intelligibility of speech in noise, suggest that it may be possible to train talkers to improve their performance with currently

available speech recognizers. Directed feedback could be provided to talkers about their articulation and how it should be selectively modified to improve recognition. If this type of feedback scheme were employed in an interactive environment, substantial gains might also be made in reducing the variability among talkers. Thus changes in a talker's speech due to high levels of masking noise, physical or psychological stress, or cognitive load could be accommodated more easily by readjustments or retuning of an adaptive system.

The present findings also suggest that the performance of current speech recognizers could be improved by incorporating specific knowledge about the detailed acoustic-phonetic changes in speech that are due to factors in the talker's physical environment such as masking noise, physical stress, and cognitive load. Some of these factors appear to introduce reliable and systematic changes in the speech waveform and, therefore, need to be studied in much greater detail in order to develop speech recognition algorithms that display robust performance over a wide variety of conditions.

ACKNOWLEDGMENTS

The research reported here was supported, in part, by Contract No. AF-F-33615-86-C-0549 from Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, Ohio, and, in part, by NIH Research Grant NS-12179. This report is Tech. Note 88-01 under the contract with AAMRL. We thank Cathy Kubaska, Howard Nusbaum, and Moshe Yuchtman for numerous contributions in carrying out this project. We also thank Diane Kewley-Port for her help in collecting the speech from the two talkers, and Michael Cluff for his help in running subjects in the perceptual experiments.

Bernacki, B. (1981). "WAVMOD: A program to modify digital waveforms," in *Research on Speech Perception Progress Report No. 7* (Speech Research Laboratory, Indiana University, Bloomington, IN), pp. 275-286.

Chen, F. R. (1980). "Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level," Unpublished Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Cooper, W. E., Eady, S. J., and Mueller, P. R. (1985). "Acoustical aspects of contrastive stress in question-answer contexts," *J. Acoust. Soc. Am.* 77, 2142-2156.

Delattre, P. (1969). "An acoustic and articulatory study of vowel reduction in four languages," *Int. Rev. Appl. Linguist.* 7, 295-325.

Doddington, G. R., and Schalk, T. B. (1981). "Speech recognition: Turning theory to practice," *IEEE Spectrum* 18, 26-32.

Draeger, G. L. (1951). "Relationships between voice variables and speech intelligibility in high level noise," *Speech Monogr.* 18, 272-278.

Dreher, J. J., and O'Neill, J. J. (1957). "Effects of ambient noise on speaker intelligibility for words and phrases," *J. Acoust. Soc. Am.* 29, 1320-1323.

Hanley, T. D., and Steer, M. D. (1949). "Effect of level of distracting noise upon speaking rate, duration and intensity," *J. Speech Hear. Disord.* 14, 363-368.

Kerstein, Z. A. (1982). "An evaluation of automatic speech recognition under three ambient noise levels," paper presented at the Workshop on Standardization for Speech I/O Technology, National Bureau of Standards, Gaithersburg, MD, 18-19 March.

Klatt, D. H. (1975). "Vowel lengthening is syntactically determined in connected discourse," *J. Phon.* 3, 129-140.

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics* (Oxford U. P., London).

Lane, H. L., and Tranel, B. (1971). "The Lombard sign and the role of hearing in speech," *J. Speech Hear. Res.* 14, 677-709.

Lane, H. L., Tranel, B., and Sisson, C. (1970). "Regulation of voice communication by sensory dynamics," *J. Acoust. Soc. Am.* 47, 618-624.

Lieberman, P. (1960). "Some acoustic correlates of word stress in American English," *J. Acoust. Soc. Am.* 32, 451-454.

Lombard, E. (1911). "Le signe de l'elevation de la voix," *Ann. Mal. Orel. Larynx* 37, 101-119. (Cited by Lane and Tranel, 1971.)

Miller, J. L. (1981). "Effects of speaking rate on segmental distinctions," in *Perspectives on the Study of Speech*, edited by P. D. Eimas and J. L. Miller (Erlbaum, Hillsdale, NJ).

Neben, G., McAulay, R. J., and Weinstein, C. J. (1983). "Experiments in isolated word recognition using noisy speech," *Proc. Int. Conf. Acoust. Speech Signal Process.*, 1156-1158.

Nusbaum, H. C., and Pisoni, D. B. (1987). "Automatic measurement of speech recognition performance: a comparison of six speaker-dependent recognition devices," *Comput. Speech Lang.* 2, 87-108.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1985). "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," *J. Speech Hear. Res.* 28, 96-103.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech Hear. Res.* 29, 434-446.

Pickett, J. M. (1956). "Effects of vocal force on the intelligibility of speech sounds," *J. Acoust. Soc. Am.* 28, 902-905.

Pollack, I., and Pickett, J. M. (1958). "Masking of speech by noise at high sound levels," *J. Acoust. Soc. Am.* 39, 127-130.

Rollins, A., and Wiesen, J. (1983). "Speech recognition and noise," *Proc. Int. Conf. Acoust. Speech Signal Process.*, 523-526.

Rostolland, D. (1982a). "Acoustic features of shouted voice," *Acustica* 50, 118-125.

Rostolland, D. (1982b). "Phonetic structure of shouted voice," *Acustica* 51, 80-89.

Rostolland, D. (1985). "Intelligibility of shouted voice," *Acustica* 57, 103-121.

Rostolland, D., and Parant, C. (1974). "Physical analysis of shouted voice," paper presented at the Eighth International Congress on Acoustics, London.

Tolhurst, G. C. (1954). "The effect on intelligibility scores of specific instructions regarding talking," Joint Project Rep. No. 35, U.S. Naval School of Aviation Medicine, Naval Air Station, Pensacola, FL, 30 November.

Tolhurst, G. C. (1955). "The effects of an instruction to be intelligible upon a speaker's intelligibility, sound pressure level and message duration," Joint Project Rep. No. 58, U.S. Naval School of Aviation Medicine, Naval Air Station, Pensacola, FL, 30 July.

Webster, J. C., and Klumpp, R. G. (1962). "Effects of ambient noise and nearby talkers on a face-to-face communication task," *J. Acoust. Soc. Am.* 34, 936-941.

An addendum to "Effects of noise on speech production: Acoustic and perceptual analyses" [J. Acoust. Soc. Am. 84, 917-928 (1988)]

W. Van Summers, Keith Johnson, David B. Pisoni, and Robert H. Bernacki
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405

(Received 16 June 1989; accepted for publication 6 July 1989)

The authors respond to Fitch's comments [H. Fitch, J. Acoust. Soc. Am. 86, 2017-2019 (1989)] on an earlier paper. New analyses are presented to address the question of whether $F1$ differences observed in the original report are an artifact of linear predictive coding (LPC) analysis techniques. Contrary to Fitch's claims, the results suggest that the $F1$ differences originally reported are, in fact, due to changes in vocal tract resonance characteristics. It is concluded that there are important acoustic-phonetic differences in speech when talkers speak in noise. These differences reflect changes in both glottal and supraglottal events that are designed to maintain speech intelligibility under adverse conditions.

PACS numbers: 43.70.Fq, 43.70.Gr

INTRODUCTION

This issue of the *Journal* contains a Letter of the Editor by Fitch (1989) that is critical of several aspects of our recent report dealing with the acoustic characteristics and intelligibility of speech produced in noise (Summers *et al.*, 1988). Fitch's primary concerns are as follows: First, our discussion should have focused more on the underlying physiological bases of the acoustic differences we reported; in particular, there was little mention of how our data on spectral tilt relate to previous studies in which changes in tilt are associated with changes in vocal effort and with specific changes in glottal waveform shape; second, we did not adequately discuss the possible relationship between our spectral tilt data and our intelligibility results; and, third, our results concerning $F1$ frequency changes in the presence of noise may have been an artifact of the LPC methods used to estimate formant frequencies and may not reflect true changes in vocal tract resonances. We address each of these criticisms, in turn, below.

Fitch's initial criticism of our paper is that our discussion of differences in the acoustic properties of Lombard versus normal speech did not address the underlying source of these differences in speech production. Because we did not collect physiological data, it was not possible to specify exactly what articulatory changes might have been made by our talkers. We agree that, in order to understand the articulatory differences between Lombard and normal speech, it is necessary to gather both articulatory and acoustic data. However, we believe that our findings represent an important contribution to understanding acoustic and perceptual consequences of speaking in noise because they show that, in addition to previously reported acoustic properties of Lombard speech (increased amplitude, $F0$, and duration), there are also reliable spectral differences between Lombard and normal speech.

According to Fitch, the changes in spectral tilt and $F1$ frequencies that we report may all be related to changes in laryngeal behavior. Concerning changes in spectral tilt, we conclude that Fitch is basically correct in her criticisms of

our work, although the issue is not as simple as she suggests. Concerning her criticisms of the $F1$ differences, we disagree more strongly. We will argue that the different average $F1$ values which we found for Lombard speech versus speech in the clear reflect real changes in supralaryngeal articulation and are not due to possible artifacts of the measurement procedures. Although we do not have articulatory data to support this claim, we are able to reject Fitch's explanation of our $F1$ data.

1. SPECTRAL TILT AND VOCAL EFFORT

Fitch begins by pointing out an important relationship between our data on spectral tilt and previous research on vocal effort. She compares our results with those of Fant (1959, 1960) and others who reported differences in spectral tilt associated with changes in effort. The changes in spectral tilt reported by Fant are similar to those reported in our paper for Lombard speech. Fant states that these differences in tilt are due to changes in the abruptness of the transitions from the closed to open (and open to closed) phase of the glottal cycle. A recent study of glottal waveforms for speech produced in a variety of environments agrees with this description for loud versus normal speech (Cummings *et al.*, 1989). However, Cummings *et al.* did not find comparable differences between Lombard speech and normal speech. Indeed, they reported that glottal waveforms for Lombard speech are more similar to those of "clear speech" than "loud speech." These results suggest that speech produced under explicit instructions to increase vocal effort may differ in a number of ways from Lombard speech. In Lombard speech, speakers apparently modify the quality of their speech to maintain intelligibility and do not simply increase vocal effort (see Lane and Tranel, 1971; Lane *et al.*, 1970). Therefore, the articulatory changes that occur as a response to increased noise in the environment may not be generalizable to all situations in which vocal effort is increased. Nevertheless, the similarity between our spectral tilt data and Fant's results is important and should have been discussed in our original article.

A related point concerns how changes in spectral tilt may influence intelligibility. Fitch cites an argument originally made by Rostolland (1982) that the decrease in spectral tilt observed in loud speech may improve intelligibility because the frequency region receiving the largest amplitude boost is the region in which hearing sensitivity is greatest. Fitch suggests that spectral tilt differences may explain why Lombard speech was more intelligible in our perceptual experiments. Once again, Fitch's point is an important one that should have been mentioned. However, while decreases in spectral tilt may be beneficial to intelligibility, it is also very likely that durational increases observed in Lombard speech also contribute substantially to increased intelligibility. These durational effects appear to be analogous to those found in "clear speech" (Chen, 1980; Picheny *et al.*, 1986), where subjects are specifically instructed to produce highly intelligible utterances. In short, while changes in spectral tilt may be one factor producing increased intelligibility for Lombard speech, it seems unlikely that these changes are the only ones relevant.

II. F_0 AND LPC ESTIMATES OF F_1

In the second part of her letter, Fitch focuses on our analyses of formant frequencies, particularly the first formant. Using LPC techniques to estimate formant frequencies, we observed increases in F_1 for utterances produced in noise. Fitch suggests that these differences may be due to the use of LPC modeling and may not actually reflect changes in vocal tract resonances. She points out two types of biases inherent in LPC modeling that may have played a role. First, LPC pole values (that are used in estimating formant values) tend to have a bias towards the nearest harmonic of the fundamental. As a result, estimates of formant frequency for a given utterance are influenced by its fundamental frequency.¹ The effect of F_0 on formant estimates is most pronounced for high F_0 values because of the wider spacing of harmonics. For this reason, the bias was of less concern in our study than if we had examined women's or children's speech. Nevertheless, Fitch points out that our data do show increases in fundamental frequency for utterances produced in noise. She argues further that the increases in F_0 may have caused the LPC-based estimates of F_1 to increase accordingly without any actual change in vocal tract resonant frequencies.

Fitch goes on to describe the LPC estimate of F_1 as "riding up with the harmonics." She seems to be suggesting here that increases in F_0 are consistently related to increases in formant peak estimates. However, an increase in F_0 does not necessarily cause harmonics near a given pole to also increase. Consider a vocal tract configuration with an F_1 frequency at 500 Hz. If F_0 is 100 Hz, there will be a harmonic that coincides precisely with the formant peak. The LPC model will presumably do an accurate job of estimating F_1 frequency in this case. However, if $F_0 = 120$, the nearest harmonic will be at 480 Hz and the LPC estimate of F_1 will "migrate" towards this value. Thus, in some circumstances, an increase in F_0 may actually lead to a decrease in the LPC estimate of F_1 . This observation calls into question the direct link that Fitch attempts to draw between our F_0 and F_1

results. While a change in F_0 may influence estimates of formant frequencies, F_0 increases will not necessarily bias estimates of formant frequency upwards (see Atal and Schroeder, 1974).

Nevertheless, some of the data reported in our original article did suggest a close relationship between F_0 and F_1 : Speaker SC showed higher mean values for both F_0 and F_1 when speaking in noise, whereas speaker MD showed fairly stable F_0 and F_1 values across quiet and noise conditions. We therefore reported an analysis of covariances (ANCOVA) on the data for SC, to examine the relationship between noise condition and F_1 , with F_0 treated as a covariate. In this analysis, the linear relationship between F_0 and F_1 was accounted for prior to examining whether F_1 frequency varied across noise conditions. The results suggested that F_1 tended to increase when speaking in noise apart from any increase in F_0 . Fitch finds this attempt to dissociate F_0 from F_1 unconvincing. She suggests that the relationship between F_0 and F_1 may not be linear since F_1 values for various utterances are not associated with the same harmonic(s) of F_0 . Her point is that a given increase in F_0 will cause a greater frequency change in higher harmonics of F_0 than in lower harmonics and therefore may have a greater effect on F_1 measurements for vowels with high F_1 frequencies than for vowels with low F_1 values (if F_1 estimates are indeed "riding up with the harmonics"). Fitting a single regression line to the (F_0 , F_1) data may therefore provide a fairly poor fit when vowels containing various F_1 frequencies are analyzed together. To test this possibility, we conducted separate analyses each of the ten vocabulary items produced by speaker SC. For each word, an analysis of covariance was conducted to examine the effect of noise condition of F_1 frequency with F_0 treated as a covariate. Within each noise condition, overall mean F_1 values across vocabulary items were then computed based on the adjusted means from these separate analyses. These adjusted F_1 values are as follows: quiet—490.4 Hz; 80-dB noise—520.5 Hz; 90-dB noise—518.4 Hz; 100-dB noise—532 Hz. These values are similar to the values from the ANCOVA originally reported in Summers *et al.* (1988). Clearly, the tendency for F_1 to increase in noise is still present.

Fitch claims that LPC-based estimates of F_1 will increase with increases in F_0 . If this is true, then there should be a positive relationship between F_0 and F_1 for tokens within a given condition as well as across conditions. We therefore examined the relationship between F_0 and F_1 across the five tokens of each word produced in each noise condition by SC. The F_0 and F_1 values for each token were used in a linear regression analysis, with F_0 used as a predictor of F_1 . Separate analyses were carried out for each vocabulary item within each noise condition for a total of 40 analyses (ten items \times four noise conditions). If F_0 increases are consistently associated with F_1 increases, the resulting regression lines should have positive slopes. Of the 40 regression lines computed, 22 showed positive slopes, while 18 showed negative slopes. Clearly, the results fail to establish any strong positive relationship between F_0 and LPC-based estimates of F_1 for tokens of an utterance produced in a given noise condition.

III. SPECTRAL TILT AND LPC ESTIMATES OF F_1

The second type of bias that Fitch suggests may have influenced our F_1 results relates to how changes in spectral tilt influence the LPC model. Changes in spectral tilt, which we observed in our study, produce differences in the relative amplitudes of harmonics near a given formant. This, in turn, could affect the LPC estimates of formant frequency without any change in vocal resonance characteristics. On initial inspection, Fitch's suggestion does not appear to be consistent with our data, since speaker MD showed decreases in spectral tilt without accompanying changes in F_1 frequency.

To further test the relationship between spectral tilt and F_1 frequency in our data, we conducted an analysis of covariance on speaker SC's F_1 data, treating noise condition and utterance as independent variables and spectral tilt² as a covariate. The adjusted mean F_1 frequencies from this analysis were: quiet—499 Hz; 80-dB noise—522.6 Hz; 90-dB noise—514.3 Hz; 100-dB noise—525.4 Hz. These results suggest that a portion of the F_1 increase observed in the data between the quiet and noise conditions is related to changes in spectral tilt; that is, the change in F_1 frequency across conditions is reduced when spectral tilt is entered as a covariate in the analysis. Nevertheless, as in the analysis of variance originally reported, the main effect of noise on F_1 frequency was statistically significant in the analysis of covariance [$F(3,159) = 4.98, p < 0.0025$]. Thus the effects of noise on F_1 for speaker SC cannot be completely accounted for by variability in spectral tilt.

In discussing the influence of tilt on F_1 measurements, Fitch compares our results with data reported by Makhoul and Wolf (1972), who examined the effect of preemphasis on formant estimates. Makhoul and Wolf report that preemphasis, which has a substantial effect on spectral tilt in the F_1 region, also affects F_1 frequency measurements.³ If the changes in spectral tilt that we observed when talkers speak in noise are similar in kind and magnitude to changes due to preemphasis, it would follow that the F_1 changes we observed might then appropriately be ascribed to changes in tilt. However, the change of spectral tilt which occurs by changing the preemphasis factor is not the same as the change of spectral tilt that results from changing the glottal source function. Energy levels at low frequencies (around F_1) are most affected by changes in preemphasis. On the other hand, in naturally occurring changes of the glottal sources spectrum, low frequencies are not affected as much as high frequencies (Rostolland, 1982).

We manipulated the spectral tilt of SC's utterances produced in quiet in order to determine how a change in spectral tilt approximating the change observed in noise would affect F_1 frequency measurements. A 51-pole filter was designed to model the difference of the long term spectra of vowels produced in the 100-dB noise condition and vowels produced in the quiet (an average difference of approximately 2 dB per octave). The items produced in the quiet were then filtered and reanalyzed. The slope of the spectrum of these "filtered-quiet" vowels was virtually identical to the slope for SC's vowels produced in 100-dB noise. If the slope of the spectrum has the kind of effect on LPC estimates of F_1 that Fitch claims, then we would expect the average F_1 of the

filtered vowels to match that of the vowels produced in noise. This was not the case. The vowels produced in noise had an average F_1 of 531 Hz, while the filtered-quiet vowels had an average F_1 of 507 Hz.⁴

We then examined how an equivalent (2 dB per octave) change in tilt due to preemphasis might influence spectral shape and estimates of F_1 frequency. In our original analysis, we used a preemphasis factor of 99% throughout. By reanalyzing the utterances produced in the 100-dB noise condition using a 66% preemphasis factor, we altered the overall spectral tilt of these utterances by approximately 2 dB per octave, giving them an average tilt approximating that of the items produced in quiet. As already mentioned, this method of altering spectral tilt has its main effect on the low-frequency portion of the spectrum. Thus it is not surprising that this manipulation had a pronounced effect on the LPC estimate of F_1 . The average estimated F_1 for the items produced in noise was 531 Hz when preemphasis was 99%, while the average F_1 of the same items with preemphasis of 66% was 494 Hz. This effect of preemphasis on estimates of F_1 replicates Makhoul and Wolf's (1972) findings.⁵ Our conclusion, then, is that not all sources of spectral tilt have comparable effects on LPC estimates of F_1 . The change in spectral tilt produced by preemphasis techniques has a much greater effect on F_1 than the change in spectral tilt produced by adjustment at the glottis.

IV. F_0 and F_1 IN PERCEPTION

We have mentioned that the LPC estimate of F_1 fluctuates around the actual F_1 as a function of F_0 (Atal and Schroeder, 1974). It is also important to note that perceived F_1 varies as a function F_0 and that this perceptual variation (for the range of F_0 variation present in our data) is very similar to the variation found in LPC estimates of F_1 . Changes in the relative amplitudes and spacing of harmonics in the F_1 region influence hearers' estimates of F_1 (Carlson *et al.*, 1975; Darwin and Gardner, 1985; Chistovich, 1985; Sundberg and Gauffin, 1978; Assmann and Nearey, 1987). Darwin and Gardner (1985) found that listeners' estimates of F_1 (as determined in a matching task) were shifted down when the amplitude of a harmonic somewhat below the actual F_1 was increased and that estimated F_1 was shifted up when the amplitude of a harmonic somewhat above the actual F_1 was increased. So, in a sense, it can be said that listeners' estimates of F_1 "ride up the harmonics," but, as with LPC estimation, the location of the harmonics relative to the actual vocal tract resonance determines whether the estimated F_1 will be lower or higher than the actual F_1 . In work that related perceptual estimation of F_1 and LPC estimation, Assman and Nearey (1987) found that LPC estimates of F_1 for tokens in which the amplitudes of harmonics in the F_1 region were manipulated correspond very closely to listeners' estimates of F_1 (in a matching task). Again, the fluctuation in LPC estimates of F_1 that have to do with changes in F_0 may have a parallel in the perceptual estimation of F_1 (Ref. 6).

V. CONCLUSION

In summary, Fitch criticizes our recent article on two counts. First, she identifies several points in our presentation in which our findings should have been related to previously published work on increased vocal effort and shouted speech. We are generally in agreement with these criticisms. The similarity between our spectral tilt data and previous data examining vocal effort should have been noted. While we have indicated certain reservations about the appropriateness of claiming that Lombard speech is the same as other forms of speech produced with increased vocal effort, the similarity between our spectral tilt data and that of Fant (1959, 1960) should have been discussed. Also, we should have discussed Rostolland's (1982) suggestions concerning the possible relationship between spectral tilt and speech intelligibility. Clearly, if Fitch had served as a reviewer of our paper, these suggestions would have been included and no doubt would have improved the discussion.

Fitch's second criticism focuses more on the scientific merit of our study than on the completeness of our discussion of its results. She points out several possible biases inherent in our use of LPC methods of estimating formant frequencies and goes on to suggest that these biases may explain our $F1$ data. As a result, Fitch contends that our findings concerning $F1$ may not reflect real changes in supralaryngeal articulation and are therefore uninteresting. We disagree with both of these criticisms. The additional analyses we report here demonstrate that the $F1$ differences we originally found were not simply due to changes in $F0$ or spectral tilt.

Characterizing the acoustic changes in speech that occur when talkers speak in noise is a research problem that has received relatively little attention in the past (however, see Stanton, 1988). Few investigations have been concerned with the consequences of these acoustics changes for speech perception and speech recognition. While information concerning articulatory behavior can be inferred from acoustic data, more direct physiological studies of speech production in severe environments are clearly needed to learn more about how talkers adjust and modify the way they speak when there are changes in their immediate acoustic environment. Most of our current knowledge about speech production and perception has been obtained with cooperative talkers in benign environments with little if any cognitive load or stress. However, most of real-world speech communication is carried out in less than the idealized conditions typically found in the laboratory. Despite the criticism raised by Fitch, we believe that our findings are valid and they suggest important acoustic-phonetic differences in speech produced in noise. To our knowledge, these findings have not been reported before in the literature in connection with Lombard speech. We believe our results are directly relevant not only to the development of new and more robust speech recognition algorithms but to our continued search for a better understanding of human speech perception as it takes place in a wide variety of environments.

ACKNOWLEDGMENTS

Preparation of this paper was supported, in part, by a contract with the Armstrong Aerospace Medical Research

Laboratory, Wright-Patterson AFB, Ohio, and, in part, by an NIH training grant to Indiana University. We thank Ken Stevens for his suggestions on a number of issues.

¹The tendency for formant estimates to be biased toward nearby harmonics of the fundamental is not limited to estimates based on LPC modeling. The tendency is also present for formant estimates based on spectrograms (Lindblom, 1962).

²The measure of spectral tilt was the slope of a regression line fit to the spectrum of the highest amplitude analysis frame of a given vowel.

³Fitch quotes a brief passage from Makhoul and Wolf (1972) regarding the influence of preemphasis on formant frequency measurements. It is interesting to note that the sentence immediately following the quoted material states: "However, these shifts are not significant in general and can be disregarded for many applications."

⁴The unfiltered quiet vowels had an average $F1$ of 488.5 Hz, so it appears that this tilt manipulation may have had some influence on estimated $F1$. However, the $F1$ values estimated by root solving are: noise—524.4; quiet—486.3; filtered quiet—490.6 Hz.

⁵The results are essentially the same for $F1$ estimated by peak picking and root solving. We report estimates made by peak picking because this was the method used in Summers *et al.* (1988).

⁶Another type of perceptual link between $F0$ and $F1$ should be mentioned. Much of the research having to do with $F0$ normalization has assumed the existence of some type of speaker normalization process in speech perception (Miller, 1953; Fujisaki and Kawashima, 1968; Slawson, 1968) rather than a single auditory effect in formant peak estimation (Trautman, 1981). Johnson (1989) has found that, when speaker identity is controlled independently, the effect of $F0$ on the perceptual evaluation of $F1$ is quite small. He suggests (following Slawson, 1968) that $F0$ plays a role in vowel identification at two levels: (1) at a psychophysical level in estimating formant peaks and (2) at a cognitive level as speaker characteristics are taken into account during vowel identification. Johnson's (1989) findings suggest that the rather large effects of $F0$ on perceived $F1$ which have been reported before are the result of the second level of processing.

Assmann, P. F., and Nearey, T. M. (1987). "Perception of front vowels: The role of harmonics in the first formant region," *J. Acoust. Soc. Am.* 81, 520-534.

Atal, B. S., and Schroeder, M. R. (1974). "Recent advances in predictive coding—Applications to speech synthesis," *Speech Communication Seminar*, Stockholm.

Carlson, R., Fant, G., and Granström, B. (1975). "Two-formant models, pitch and vowel perception," an *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London).

Chen, F. R. (1980). "Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level," unpublished master's thesis, MIT, Cambridge, MA.

Chistovich, L. A. (1985). "Central auditory processing of peripheral vowel spectra," *J. Acoust. Soc. Am.* 77, 798-805.

Cummings, K. E., Clements, M. A., and Hansen, J. H. L. (1989). "Estimation and comparison of the glottal source waveform across stress styles using glottal inverse filtering," *IEEE Proc., Southeastcon.*, April, 1989, pp. 776-781.

Darwin, C. J., and Gardner, R. B. (1985). "Which harmonics contribute to the estimation of first formant frequency?," *Speech Commun.* 4, 231-235.

Fant, G. (1959). "Acoustic analysis and synthesis of speech with applications to Swedish," *Ericsson Tech.* 15, 1-106.

Fant, C. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fitch, H. (1989). "Comments on effects of noise on speech production: Acoustic and perceptual analyses" [*J. Acoust. Soc. Am.* 84, 917-928 (1988)], *J. Acoust. Soc. Am.* 86, 2017-2019.

Fujisaki, H., and Kawashima, T. (1968). "The role of pitch and higher formants in the perception of vowels," *IEEE Trans. Audio and Electroacoust.* 16, (1), 73-77.

Johnson, K. (1989). "Intonational context and $F0$ normalization," *J. Acoust. Soc. Am.* (submitted).

Lane, H. L., and Tranel, B. (1971). "The Lombard sign and the role of hearing in speech," *J. Speech and Hear. Res.* 14, 677-709.

Lane, H. L., Tranel, B., and Sisson, C. (1970). "Regulation of voice com-

- munication by sensory dynamics," *J. Acoust. Soc. Am.* **47**, 618-624.
- Lindblom, B. (1962). "Accuracy and limitations of sonagraph measurements," in *Proc. Fourth Int. Congr. Phon. Sci.*, Helsinki, 188-202.
- Makhoul, J. I., and Wolf, J. J. (1972). "Linear prediction and the spectral analysis of speech," Bolt Beranek and Newman Inc., Cambridge, MA, NTIS AD-479066, Rep. 2304.
- Miller, R. L. (1953). "Auditory tests with synthetic vowels," *J. Acoust. Soc. Am.* **25**, 114-121.
- Picheny, M. A., Durlach, N. I., and Braid, L. D. (1986). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," *J. Speech Hear. Res.* **29**, 434-446.
- Rostolland, D. (1982). "Acoustic features of shouted voice," *Acustica* **50**, 118-125.
- Slawson, A. W. (1968). "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency," *J. Acoust. Soc. Am.* **43**, 87-101.
- Staton, B. J. (1988). "Robust recognition of loud and Lombard speech in the fighter cockpit environment," unpublished Ph.D. dissertation, Purdue University.
- Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (1988). "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Am.* **84**, 917-928.
- Sundberg, J., and Gauffin, J. (1978). "Waveform and spectrum of the glottal source," *STL-QPSR* **2-3**, 35-50.
- Trautmüller, H. (1981). "Perceptual dimension of openness in vowels," *J. Acoust. Soc. Am.* **69**, 1465-1475.

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 15 (1989)

Indiana University

Effects of Cognitive Workload on Speech Production: Acoustic Analyses¹

W. Van Summers², David B. Pisoni, Robert H. Bernacki

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405*

¹This work was supported by a contract from Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, Contract No. AF-F-33615-86-C-0549 to Indiana University, Bloomington, Indiana.

²Now at Army Audiology and Speech Center, Walter Reed Army Center, Washington, DC 20307.

Abstract

The work presented here examined the effects of cognitive workload on speech production. Cognitive workload was manipulated by having subjects perform a visual compensatory tracking task while they were speaking test sentences. Sentences produced in this workload condition were compared with control sentences which were produced in a no workload condition. Subjects produced utterances in the workload condition with increased amplitude, increased amplitude variability, decreased spectral tilt, increased F0 variability, and increased speaking rate. These changes involve both laryngeal and sublaryngeal structures and changes in articulatory timing. There was no evidence of vowel reduction or other changes in subjects' abilities to achieve supralaryngeal articulatory targets.

Effects of Cognitive Workload on Speech Production: Acoustic Analyses

Attentional and cognitive demands placed upon pilots, flight controllers, and others involved in information-intensive jobs may influence the acoustic characteristics of their speech in demanding situations. Very little research has explored whether consistent changes can be identified in the characteristics of utterances produced in demanding or "high-workload" environments. This type of research could have several important applications. First, if speech characteristics could be identified which correlate with the level of workload an operator is experiencing, this information could be used in training and selecting operators or in testing environments for their human-factors acceptability. This information could also be important in the design of speech-recognition devices which may operate in high-workload settings. These devices must be able to tolerate changes in acoustic characteristics that occur as a result of variability in workload. The present study was aimed at exploring whether consistent changes in speech could be identified which were the result of changes in the attentional and cognitive demands of the environment.

Previous research involving workload tasks have generally assumed that workload increases are associated with increased psychological stress (e.g., Hecker, Stevens, von Bismark and Williams, 1968; Tolkmitt and Scherer, 1986). Therefore, the results of these studies have often been equated with studies in which stress is manipulated through exposure to aversive stimulation, instructions requiring subjects to lie to the experimenter (or an accomplice) or other means of increasing emotional stress (Scherer, 1979). In addition, the majority of previous studies concerned with these issues have focused on fundamental frequency (F0) characteristics as an indicator of workload or stress (see, for example, Williams and Stevens, 1969; Kuroda, Fujiwara, Okamura and Utsuki, 1976; Tolkmitt and Scherer, 1986). Few studies have examined a larger array of acoustic characteristics in an effort to produce a more complete description of the effects of increased workload on the speech signal (but see Hansen, 1988). The present study examined how changes in cognitive workload affect various acoustic characteristics of the speech signal and whether changes that can be associated with increased workload are similar to changes that have previously been ascribed to increased emotional stress.

In the present study, cognitive workload was manipulated by requiring speakers to perform an attention-demanding secondary task while speaking. The task chosen was a compensatory tracking task which was first described in Jex, McDonnell and Phatak (1966). The tracking task will be referred to as the "JEX" task hereafter. The task involved manipulating a joystick in order to keep a pointer centered between two boundaries on a computer screen (see Figure 1). The program deflected the pointer away from the center position and the subject was required to continuously compensate for the movement of the pointer by manipulating the joystick in order to keep it from crashing into one of the boundaries. Phrases which the subjects were required to produce were visually presented on the computer screen while the subjects continued to perform the JEX task.

Insert Figure 1 about here

Method

Subjects. Five male native English speakers were recruited as subjects. Three subjects (ME, TG, and EG) were psychology graduates student who were paid for their participation. Two subjects (MC and SL) were members of the laboratory and participated as part of their routine duties. All speakers were naive to the purpose of the study. None of the speakers reported a hearing or speech problem at the time of testing.

Procedure. Subjects were run individually in a single-walled sound-attenuated booth (IAC Model 401A). The subject was seated comfortably facing a video screen which contained the JEX task display and (during sessions in which speech was collected) the phrases to be produced. The subject wore a headset fitted with an Electrovoice condenser microphone (Model C090) which was attached to the headset with an adjustable boom. Once adjusted, the microphone remained at a fixed distance of 4 inches from the subject's lips throughout the experiment. Subjects wore the headset during training sessions on the JEX task and during the experimental sessions in which utterances were collected.

Subjects were trained on the JEX task alone for several days. When a subject was able to consistently perform the task at a fairly high level of difficulty, the actual experiment was conducted. During the experiment, subjects simultaneously performed the JEX task and produced phrases presented on the video screen. Test phrases consisted of /h/-vowel-/d/ utterances in the sentence frame: "Say hVd again". The English vowels, /i, e, a, æ, ɔ, u, ʌ, oʊ, ɛ/ appeared in the hVd context. During the actual experiment, JEX task difficulty was set to 80% of the level attained by the subject during training. Test phrases were only presented while the subject was performing the JEX task at this level of difficulty.

For each subject, utterances were collected in two experimental sessions. Each session consisted of 4 blocks of 20 trials with each of the 10 phrases presented twice within each block. Blocks of trials alternated between "JEX" and "control" blocks. During JEX blocks, subjects performed the JEX task while producing the test phrases; during control blocks, subjects produced the test utterances without performing any simultaneous task. For each subject, a total of 8 tokens of each phrase were produced in each condition ¹

¹For subject EG, we report data from experimental session 2 only. The level of JEX task difficulty (80% of the level attained during training) used in experimental session 1 was apparently too low for this subject. He performed the task without any crashes for the entire session and did not appear to be under any workload. JEX task difficulty was increased in session 2. With this increase in difficulty, performance on the JEX task was similar to the performance of the other subjects. The exclusion of session 1 data for EG left 4 tokens of each utterance available from each condition.

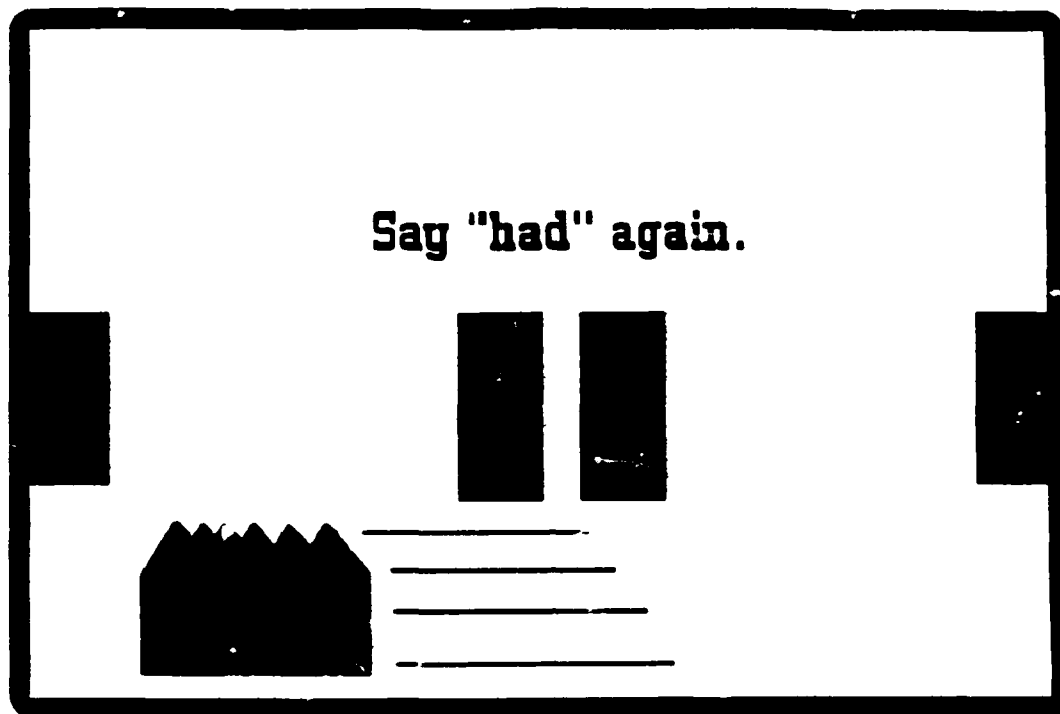


Figure 1. Illustration of the JEX compensatory tracking procedure. The subjects' task is to keep the moving pointer located at the bottom of the display from crashing into the sides of the display. During speech-collection intervals, sentences were presented in the top portion of the display as shown.

Speech Signal Processing. Test utterances were analyzed using digital signal processing techniques. Utterances were sampled and digitized on-line by a VAX 11/750 computer during the experimental sessions. Utterances were first low-pass filtered at 4.8 kHz and then sampled at a rate of 10 kHz using a 16 bit A/D converter (Digital Sound Corporation Model 2000). Each utterance was sampled into a separate waveform file.

Linear predictive coding (LPC) analysis was performed on each waveform file. LPC coefficients were calculated every 12.8 ms using the autocorrelation method with a 25.6 ms Hamming window. Fourteen linear prediction coefficients were used in the LPC analyses. The LPC coefficients were then used to calculate the short-term spectrum and overall power level of each analysis frame (window). Formant frequencies, bandwidths, and amplitudes were also calculated for each frame from the LPC coefficients. In addition, a pitch extraction algorithm was employed to determine if a given frame was voiced or voiceless and, for voiced frames, to estimate the fundamental frequency (F0).

Total duration for each phrase was determined by visual inspection and measurement from a CRT display which simultaneously presented the waveform along with time-aligned, frame-by-frame plots of amplitude, F0 (for voiced frames), and formant parameters. Cursor controls were used to locate the onset and offset of each utterance. The onset and offset of the /h/ frication, vowel, and /d/ closure segments from the hVd portion of each utterance were also identified and labelled. Following identification of utterance and segment boundaries, a program stored durational and RMS energy information for each utterance and segment. Fundamental frequency and formant frequency information were also stored for the phrase and for the vowel of the hVd portion of the phrase.

Results and Discussion

The influence of cognitive workload on various acoustic characteristics of the test utterances is described below. In each case, an analysis of variance was used to determine whether workload had a significant effect on a given acoustic measure. Separate analyses were carried out for each speaker. The analyses used phrase and workload condition (JEX or control) as independent variables and a *p* value of .05 as the critical value in all tests of statistical significance. The presentation of results will focus on the effect of workload on the various acoustic measures. The phrase variable will be discussed only in cases where a significant phrase X workload interaction was observed.

Amplitude. The upper panel of Figure 2 shows amplitudes averaged across entire phrases for utterances from the JEX and control conditions. The data are plotted separately by speaker. The lower panel of the figure shows amplitudes at the segmental level. Amplitudes of the /h/ frication, vowel, and /d/ closure portion of each hVd utterance are shown for each workload condition. An asterisk above a pair of bars indicates a significant difference between values in the JEX and control conditions for a particular speaker.

Insert Figure 2 about here

As the figure shows, there was a tendency for amplitude to increase in the JEX condition. The pattern is very consistent for speakers SL, ME, and TG who showed significantly higher amplitudes in the JEX condition for the entire phrase and for the separate /h/, vowel, and /d/ closure segments. Speaker MC showed significantly higher amplitudes in the JEX condition for the vowel and /d/ closure segments. Speaker EG did not show as clear a pattern of amplitude increases under workload as the other speakers. For this speaker, /h/ frication amplitude was significantly higher in the control condition than in the JEX condition. However, this significant main effect was mediated by a significant phrase X workload condition interaction. For EG, /h/ frication amplitude was higher in the control condition in 7 of the 10 vowel contexts. Of all of the analyses reported in this study, this was the only case of a significant phrase X workload interaction.

Amplitude variability (across utterances). Along with an increase in mean amplitude, amplitude variability from one utterance to the next also tended to increase in the workload condition. Figure 3 shows standard deviations of phrase amplitude across utterances for each condition. For the entire phrase, four of the five subjects showed an increase in amplitude variability in the JEX condition. For three of these subjects, the increase in variability was statistically significant. One subject showed the opposite pattern with significantly less amplitude variability in the JEX condition. As the lower panel of the figure shows, the pattern just described for the entire phrase was also true for amplitude variability of vowels and /d/ closure segments in the h-vowel-d context. In each case, four of the five subjects showed greater amplitude variability when performing the workload task. The effect was statistically significant for three of these four speakers in the case of vowels but was only significant for one speaker in the case of /d/ closure.

Insert Figure 3 about here

Spectral tilt. Amplitude increases are often correlated with changes in "spectral tilt". That is, high amplitude utterances generally show flatter spectra with relatively more high frequency energy than is seen in low amplitude utterances. We examined the long-term spectra of the hVd vowels in each condition to determine if the amplitude increases seen in the JEX task were correlated with decreased spectral tilt. Figure 4 shows the difference in energy between JEX-condition and control-condition vowels across 40-Hz linear frequency bands.

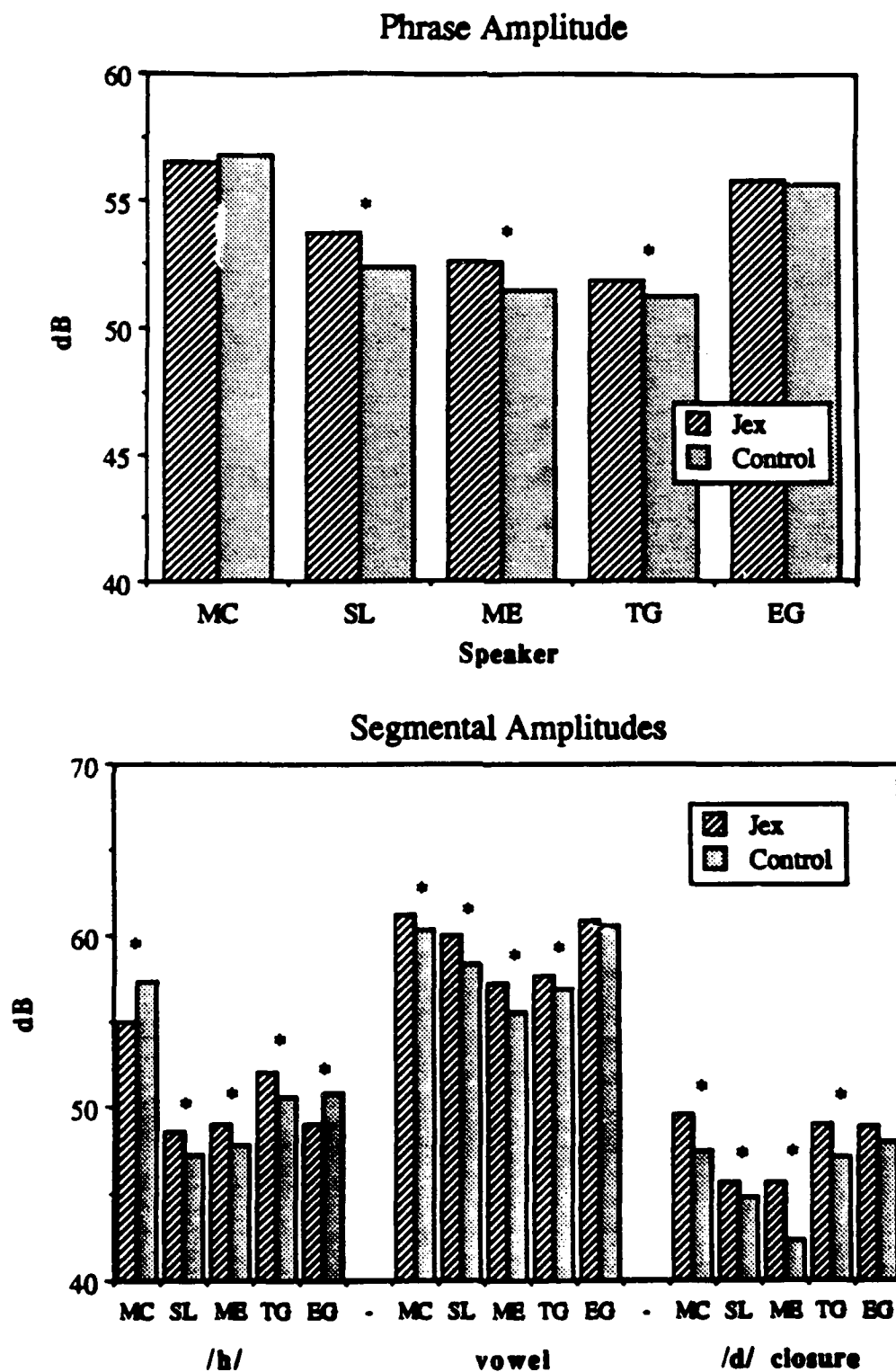


Figure 2. Phrase amplitude (upper pannel) and segmental amplitudes for "Say hVd again" utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

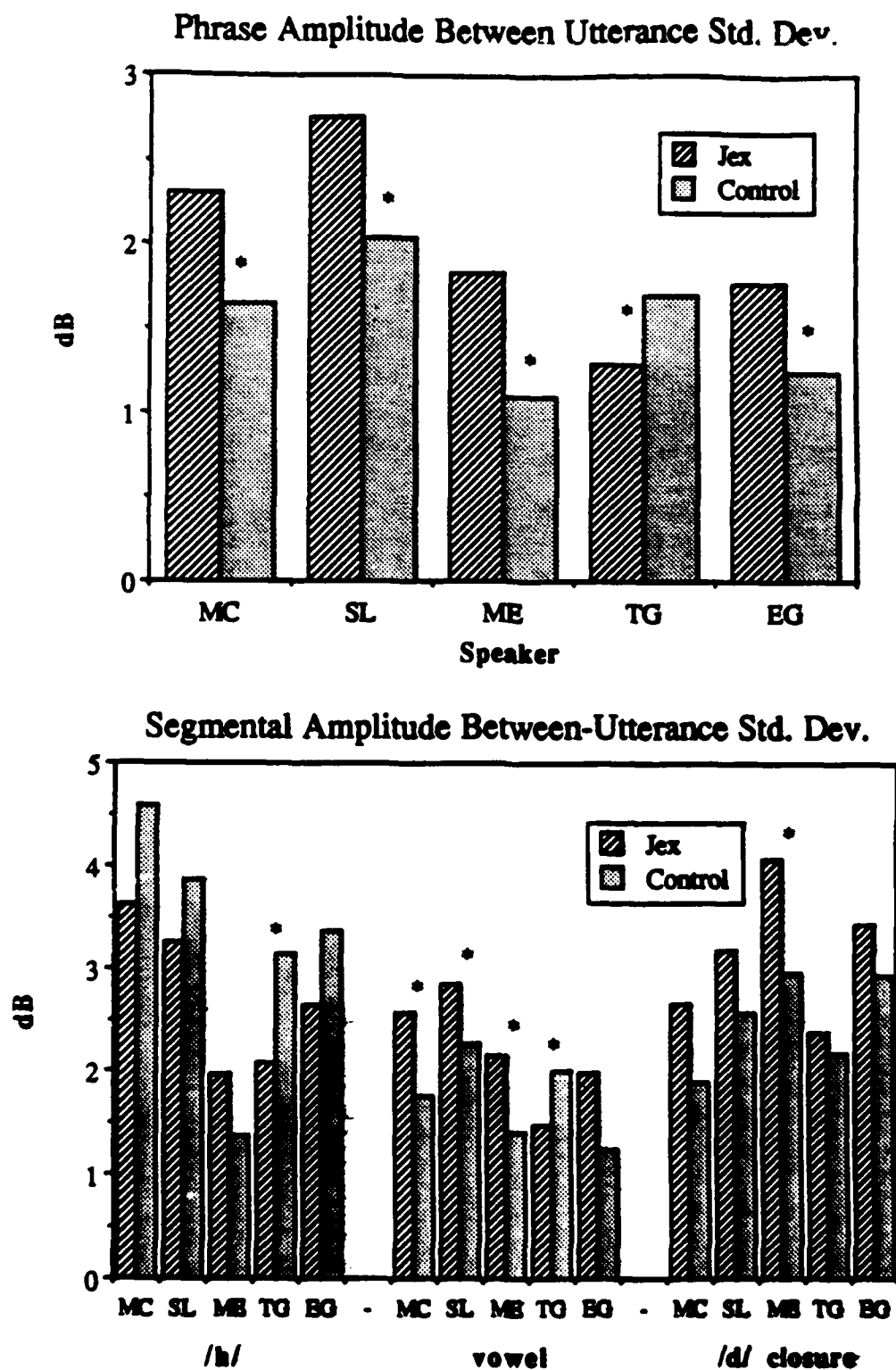


Figure 3. Between-utterance standard deviations for phrase amplitudes (upper panel) and segmental amplitudes. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker

Insert Figure 4 about here

A positive slope in these figures indicates that the difference in energy between JEX-condition and control-condition vowels is increasing with frequency. This is the pattern seen for speakers MC, SL, ME, and EG. Thus, as in the amplitude data, four of the five subjects show a consistent pattern. However, the four subjects who showed changes in spectral tilt across conditions are not the same four who showed amplitude differences. Subject EG, the subject who did not show significant amplitude differences across workload conditions, shows one of the clearest cases of changes in spectral tilt. Overall, the data suggest that the workload task produces effects on spectral tilt that are not always correlated with changes in overall amplitude.

Hansen (1988) has provided further evidence that decreases in spectral tilt under workload are not necessarily linked to amplitude increases. For the both the JEX task and the dual task examined by Hansen, spectral tilt decreased under workload without any amplitude increase.

Fundamental frequency. We also analyzed fundamental frequency for each phrase and for the hVd vowel segments in each condition. Figure 5 shows mean F0 values for the phrase and hVd vowels in each condition for each talker. Two speakers (MC and SL) showed a significant increase in F0 for the entire phrase and for the hVd vowel when performing the JEX task. The pattern of F0 increase under workload was not replicated for either the phrase or hVd vowel in the other three subjects' data.

Insert Figure 5 about here

The absence of a consistent effect of workload on mean F0 values was also reported by Hansen (1988). Hansen examined speech produced while performing the JEX task and a "dual task" requiring the simultaneous performance of two tracking tasks while speaking. Neither of these workload tasks had any consistent effect on mean F0. Conversely, previous research examining the effects of emotional stress on speech have generally reported increases in mean F0 accompanying increased stress (Williams and Stevens, 1969; Kuroda et al., 1976; Streeter, MacDonald, Apple, Krause and Galotti, 1983).

Fundamental frequency variability (within utterances). A different characteristic of the F0 data does, however, show a fairly consistent pattern across workload conditions. Figure 6 shows the standard deviations of the frame-by-frame F0 values for each phrase and for each hVd vowel in each condition. As the figure shows, three subjects showed a significant

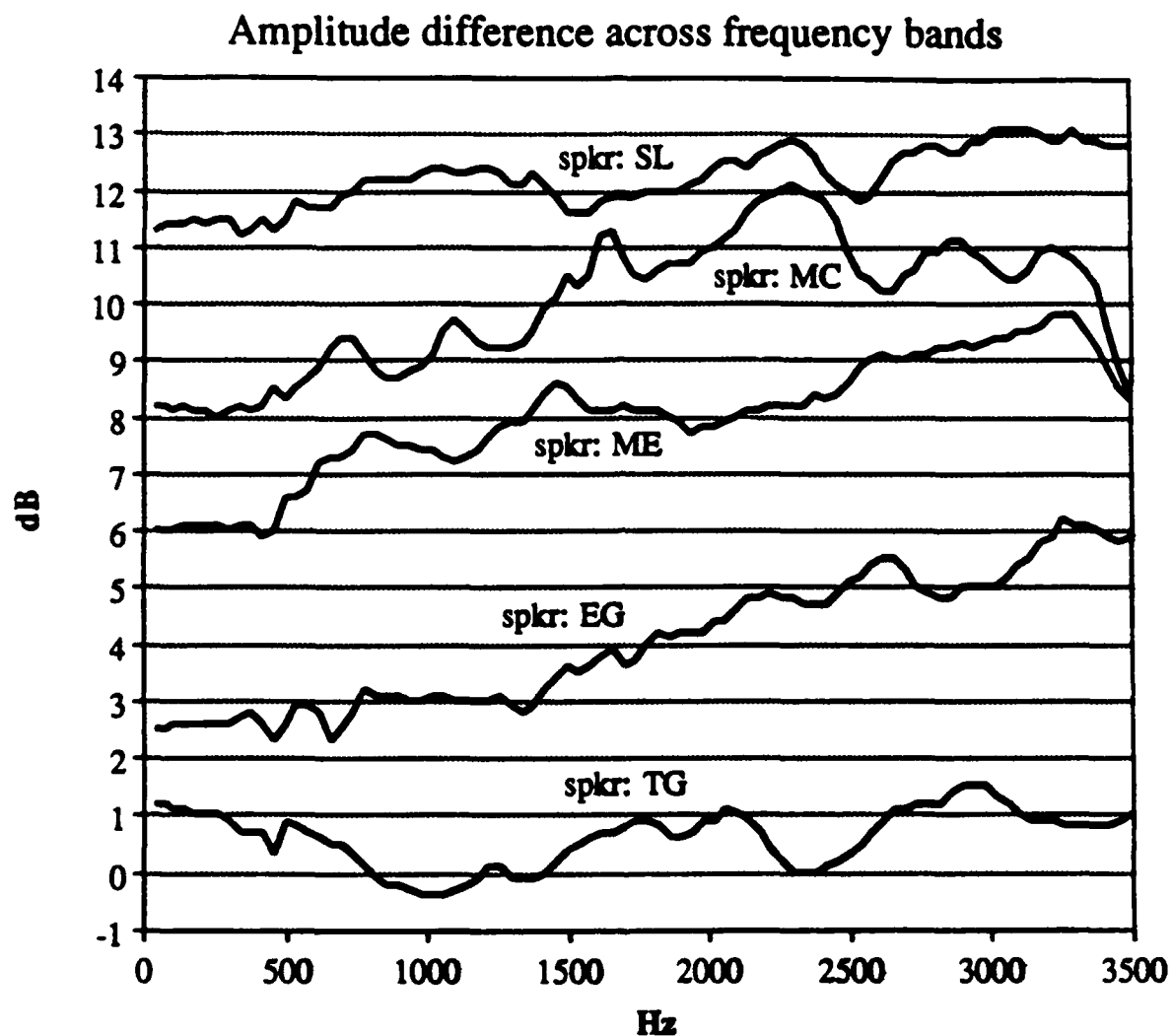


Figure 4. Mean difference in energy between utterances produced in the JEX and control conditions across frequency bands. Values are collapsed across utterances and presented separately for each speaker. For clarity of presentation, the traces for speakers EG, ME, MC and SL have been elevated by 2.5, 5, 7.5 and 10 dB respectively.

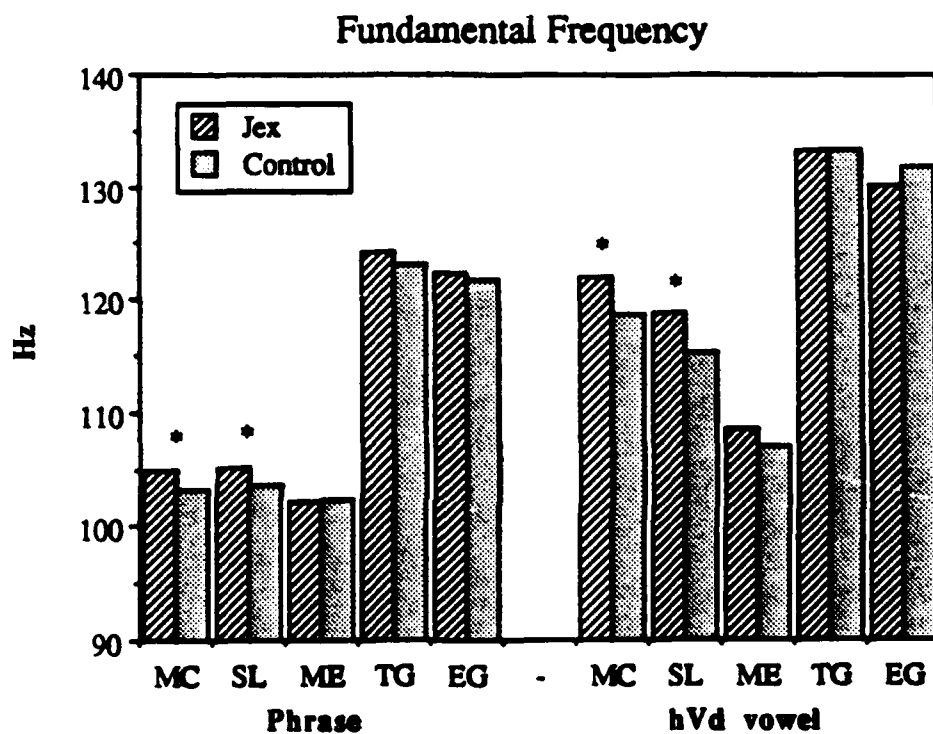


Figure 5. Mean fundamental frequency values for utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

reduction in F0 variability when performing the JEX task. A fourth subject showed the same general pattern, although the difference was not significant. The pattern is not seen in the vowel data. Given that F0 variability decreases for the entire phrase but not the vowel, the pattern is more likely due to a flattening of the overall F0 contour rather than a decrease in period-to-period F0 variability or "vocal jitter." In other words, the whole phrase is apparently produced using a monotone pitch when the subject is under workload.

Insert Figure 6 about here

The decrease in F0 variability under workload is interesting in light of previous research examining the performance of the Psychological Stress Evaluator (PSE), a commercially-available "vocal lie detector". The PSE responds to an 8-14 Hz frequency modulation (FM) in the vocal signal. Brenner, Branscomb and Schwarz (1979) report that this frequency modulation is decreased when subjects are required to perform a speeded arithmetic task. This type of decrease in F0 modulation could produce the decrease in F0 variability reported above although it should be observed for individual vowels even more readily than for entire phrases.

Duration. Figure 7 shows the effect of cognitive workload on phrase durations and segmental durations. Four of the five speakers showed significantly shorter overall phrase durations while performing the JEX task. One speaker showed the opposite pattern with longer phrase durations while under workload. This speaker showed the smallest change in phrase duration across conditions. Segmental durations also tended to be reduced while performing the JEX task. The four speakers who showed shorter phrase durations in the JEX condition also tended to show shorter /h/ frication durations and shorter /d/ closure durations. Vowel duration (in hVd words) was less consistently affected by the workload condition. The durational shortening observed for the entire phrase, the /h/ frication, and the /d/ closure, replicate results mentioned briefly in Hecker et al. (1968).

Insert Figure 7 about here

Given that the vowel in the hVd contexts was the only part of the phrase containing "new" information from trial-to-trial, speakers may have treated the production of this vowel as more important than the production of surrounding context. This may explain why vowel duration was not consistently reduced in the JEX condition while other segmental durations were reduced in the remainder of the utterance.

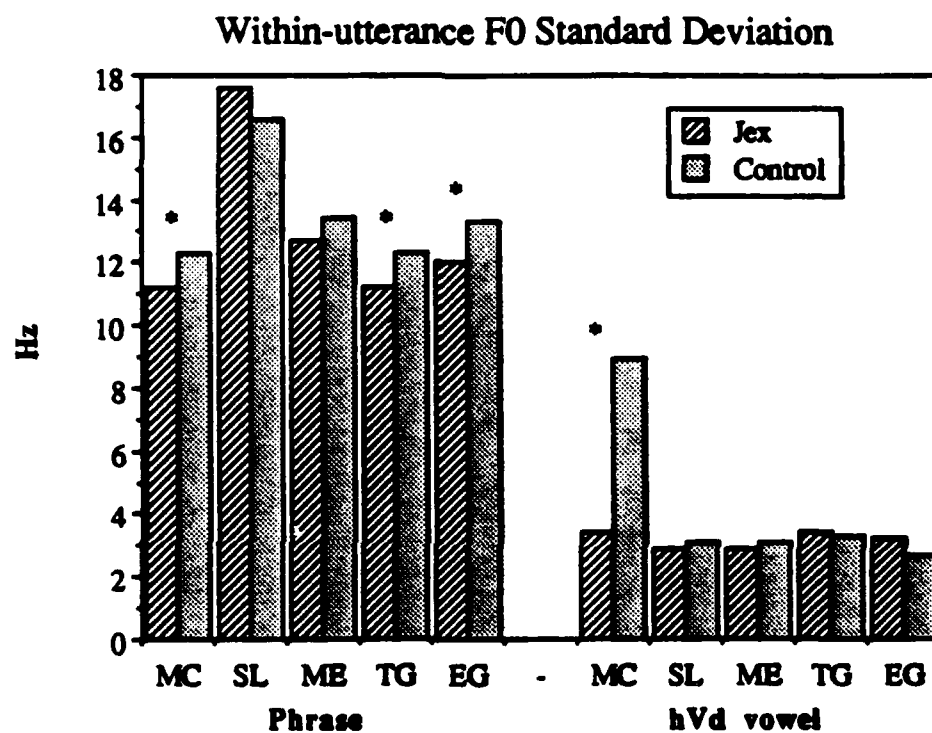


Figure 6. Mean within-utterance F0 standard deviations for utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

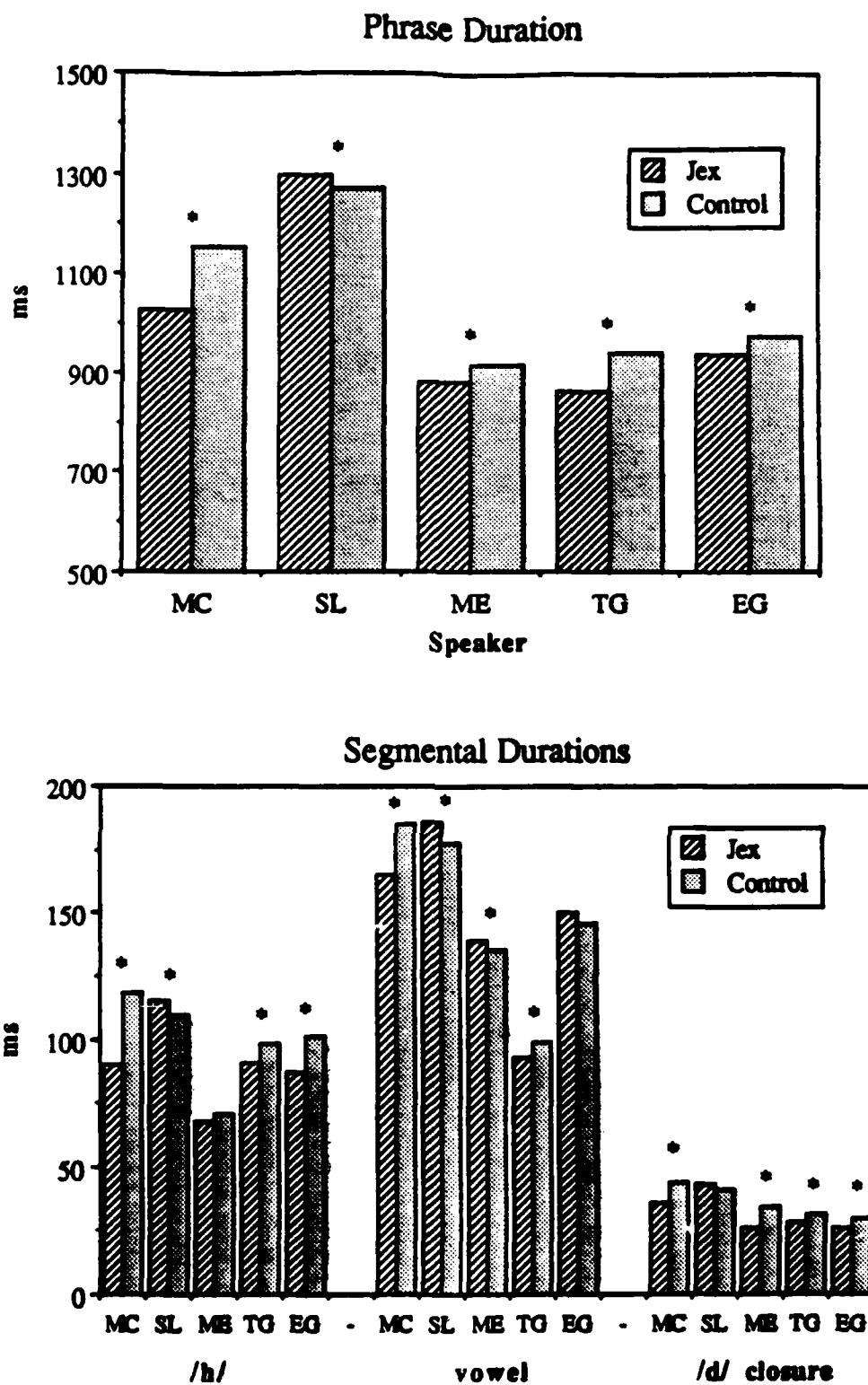


Figure 7. Mean phrase duration (upper pannel) and mean segmental duration values for utterances produced in JEX and control conditions. The * symbol appears between mean values that are significantly different. Values are collapsed across utterances and presented separately for each speaker.

Goldman-Eisler (1968) reports data on speech pauses and spontaneity that may be related to the present findings. Her data suggest that hesitations decrease and fluency increases as subjects repeat a given passage. This may explain why subjects increase their speech rate (i.e., decreased durations) for the part of the phrase which they continued to repeat. Presumably this decrease in duration would allow more time to be allocated to the workload task. On the other hand, the identity of the hVd vowel changed from trial-to-trial so that this increase in fluency/rate did not occur for this portion of the phrase.

Formant frequencies. Workload did not have a clear influence on the frequencies or bandwidths of the first three formants for any of the five speakers. Thus, it appears that workload had a greater influence on sub-laryngeal and laryngeal (source-related) functions and speech timing than it did on the supralaryngeal control of speech.

Summary and Conclusions

Very little previous research has attempted to identify consistent changes that occur in the acoustic-phonetic properties of speech produced in severe environments. Research in this area may have important implications for human-to-human and human-to-machine speech communication in demanding environments such as cockpits and air traffic control towers.

The present results show that increased cognitive workload produces a number of effects on the acoustic-phonetic properties of speech. Utterances produced under cognitive workload show higher amplitudes and greater amplitude variability between utterances. Spectral tilt was reduced for vowels produced under workload and this change in tilt was not always correlated with a change in amplitude. F0 variability within an utterance was reduced under workload, suggesting that these utterances were produced with a flatter and perhaps less expressive F0 contour. Overall phrase durations and segmental durations were also reduced under workload, suggesting an overall increase in speaking rate as workload increased.

The patterns reported here are tendencies that emerged across a small number of subjects. Some differences were not always present for each subject. We believe that these patterns may be more consistent in an actual "high workload" environment than could be seen in this investigation in which performing poorly on the workload task had only minor consequences (compare, for example, Williams & Stevens, 1968, analysis of F0 characteristics in tape-recordings of actual conversations between pilots and flight controllers versus Hecker et al.'s (1969) analysis of F0 in a laboratory task designed to increase workload). Of the five speakers examined here, subject MC performed the JEX task at the highest level of difficulty and may have been the most highly motivated of the five subjects. It is interesting to note that, in general, MC showed the most consistent effects of workload on the acoustic-phonetic properties of speech.

The absence of any effect of workload on formant frequencies in combination with the other findings suggests that the main effect of workload occurred at or below the level of the

larynx. The changes in amplitude, F0 characteristics, and spectral tilt that we found in this study may be related to changes in the shape and variability of the glottal waveform (Hecker et al, 1968). We are currently analyzing our data further to determine the extent to which the various acoustic changes described above can be ascribed to this one source.

In summary, the results of this study demonstrate a number of reliable changes in the acoustic-phonetic properties of speech produced under increased cognitive workload. The findings add to a growing body of literature showing that talkers will consistently modify their speech in response to both physical and mental changes in their immediate environments. These results have important implications for the use of speech recognition devices in severe environments.

References

- Brenner, M., Branscomb, H. H., and Schwarz G. E. (1979). Psychological stress evaluator- Two tests of a vocal measure. *Psychophysiology*, 16, 351-357.
- Goldman Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.
- Hansen, J. H. L. (1988). Analysis and compensation of stressed and noisy speech with application to robust automatic recognition. Unpublished doctoral dissertation. Georgia Institute of Technology.
- Hecker, M. H. L., Stevens, K. N., von Bismarck, G., and Williams, C. E. (1970). Manifestations of task-induced stress in the acoustic speech signal. *Journal of the Acoustical Society of America*, 44, 993-1001.
- Kuroda, I., Fujiwara, O., Okamura, N., and Utsuki, N. (1976). Method for determining pilot stress through analysis of voice communication *Aviation, Space, and Environmental Medicine*, 47, 528-533.
- Jex, H. R., McDonnell, J. D., and Phatak, A. V. (1966). A 'critical' tracking task for manual control research. In Moray, N., editor, *Mental Workload*. New York: Plenum Press.
- Scherer, K. R. (1979). Nonlinguistic vocal indicators of emotion and psychopathology. In Valdman, A., editor, *Emotions in personality and psychopathology*, pp. 493-529. New York: Academic Press.
- Streeter, L. A., MacDonald, N. H., Apple, W., Krause, R. M., and Galotti, K. M. (1983). Acoustic and perceptual indicators of emotional stress. *Journal of the Acoustical Society of America*, 73, 1354-1360.
- Tolkmitt, F. J., and Scherer, K. R. (1986) Effect of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology*, 12, 302-312.
- Williams, C. E., and Stevens, K. N. (1969) On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Medicine*, 40, 1369-1372.